



**UNIVERSITÉ
CAEN
NORMANDIE**

UNIVERSITÉ DE CAEN NORMANDIE

MASTER 2 STATISTIQUES APPLIQUÉES ET
ANALYSE DÉCISIONNELLE

RAPPORT DE PROJET

Analyse de données

ANALYSE STATISTIQUE DE L'INSERTION
PROFESSIONNELLE DES DIPLÔMÉS DE MASTER

Etudiants: *REMESHA, TANGUY, REKIK, LAMAH*
Professeur référent: *Christophe CHESNEAU*

Durée : du 05/01/ au 19/02/24

16 février 2024

Table des matières

1	Introduction	2
1.1	Contexte du projet	2
1.2	Source des données	2
1.3	Objectifs de l'étude	3
1.4	Méthodologie	3
1.4.1	Analyses descriptives	4
1.4.2	Analyses bivariées et multivariées	4
1.4.3	Modélisation statistique	5
1.4.4	Tests statistiques	5
1.4.5	Clustering et classification	6
2	Analyse descriptive	7
2.1	Au coeur des disparités salariales : Ce que les salaires médians nous révèlent	7
2.1.1	Quels facteurs sous-tendent les différences de salaires médians par domaine d'études ?	7
2.1.2	Quels mystères se dissimulent derrière les inégalités de salaires entre les sexes ?	11
2.1.3	Quels domaines d'études exacerbent le plus les inégalités salariales entre hommes et femmes ?	13
2.1.4	Les différences de salaires médians par domaine : sont-elles statistiquement significatives ?	14
2.2	Sur les chemins de l'emploi : une analyse du taux d'insertion professionnelle	16
2.2.1	Évolution des taux d'insertion professionnelle par sexe : Analyse des tendances et des disparités sur de 8 ans"	16
2.2.2	Plongée dans l'insertion professionnelle par domaine : révélations sur les tendances évolutives et les disparités cachées	17
2.3	Les emplois stables : Une réalité ou un mirage ?	20
2.3.1	Domaines offrant les emplois les plus stables : analyse des tendances récentes	20

2.3.2	La stabilité de l'emploi : une disparité persistante entre les sexes	22
2.4	Analyse bivariée des disparités de sexe	23
2.5	Corrélations entre taux d'insertion, type d'Emploi, et sexe . . .	24
3	Analyse Avancée	28
3.1	Analyse en Composantes Principales (ACP)	28
3.2	Prétraitement des Données et Gestion des Valeurs Manquantes .	28
3.3	Préparation des données et exécution de l'ACP	29
3.4	Résultats de l'Analyse en Composantes Principales	29
3.4.1	Cercle des corrélations	29
3.4.2	Interprétation des axes principaux	30
3.4.3	Interprétation des variables supplémentaires	31
3.4.4	Graphique des individus de l'ACP	33
3.5	Résultats du clustering	33
3.6	Analyse de Régression	36
3.6.1	Résultats du modèle de régression	36
3.6.2	Interprétation des effets des variables	37
3.6.3	Analyse des effets des variables sur le salaire net médian	38
3.7	Vérification des hypothèses de la régression linéaire	39
3.7.1	Diagnostic du modèle de régression	39
3.7.2	Vérification de l'indépendance des résidus	40
3.7.3	Test d'hétéroscédasticité	40
3.7.4	Normalité des résidus	42
3.7.5	Analyse de la multicollinéarité	43
3.7.6	Gestion des observations influentes	43
3.7.7	Comparaison des modèles avant et après correction des valeurs influentes	44
3.7.8	Analyse des hypothèses statistiques après correction . . .	45
3.8	Modélisation par Arbre de Régression	45
3.8.1	Préparation des données pour l'arbre de régression . . .	46
3.8.2	Construction de l'Arbre de Régression	47
4	Conclusion	49
	Appendices	53
.1	Dictionnaire des variables	54

Table des figures

2.1	Salaire Médian par Domaine	8
2.2	Détail du Salaire Médian par Domaine	9
2.3	Détail du Salaire Médian par Domaine	9
2.4	Évolution de l'écart de salaire net moyen entre les sexes de 2013 à 2020	12
2.5	Évolution de l'écart de salaire médian net entre les sexes de 2013 à 2020	14
2.6	Analyse des différences de salaire médian par domaine avec test de Kruskal-Wallis	15
2.7	Taux d'insertion par sexe au fil des années	16
2.8	Taux d'insertion par domaine au fil des années	18
2.9	Distribution des taux d'insertion par domaine	19
2.10	Taux d'emplois stables par domaine au fil des années	21
2.11	Taux d'emplois stables par sexe au fil des années	22
2.12	Répartition des genres par domaine et distribution des emplois à temps plein et des salaires nets médians	23
2.13	Matrice de corrélations entre taux d'insertion, type d'emploi, et genre	25
3.1	Cercle des corrélations des variables de l'ACP.	30
3.2	Graphique de projection des variables supplémentaires sur les axes de l'ACP.	31
3.3	Statistiques des variables supplémentaires dans l'ACP.	32
3.4	Graphique des individus.	33
3.5	Représentation des individus sur le plan factoriel avec la seg- mentation en clusters.	34
3.6	Dendrogramme résultant du clustering hiérarchique.	35
3.7	Dendrogramme 3D résultant du clustering hiérarchique.	35
3.8	Effets estimés des variables explicatives sur le salaire net médian pour les modèles complet et réduit.	37

3.9	Effets estimés des variables explicatives sur le log du salaire net médian. Les graphiques détaillent les contributions marginales des variables de genre, de statut d'emploi permanent, de statut d'emploi temporaire et de statut d'emploi indépendant, révélant les tendances et les significativités associées.	38
3.10	Graphique des résidus par rapport aux valeurs ajustées pour le modèle de régression linéaire.	39
3.11	Fonction d'autocorrélation des résidus du modèle de régression linéaire ajusté.	40
3.12	Graphique d'influence indiquant les observations potentiellement influentes. Les cercles représentent la distance de Cook, avec un intérêt particulier pour les observations dont la distance de Cook dépasse 1, ce qui indique une influence substantielle sur le modèle.	41
3.13	Graphique Q-Q des résidus standardisés du modèle. Ce graphique compare la distribution des résidus avec une distribution normale théorique. Les points s'écartant de la ligne représentent des déviations par rapport à la normalité.	42
3.14	Comparaison des coefficients estimés du modèle de régression avant et après correction des valeurs influentes. Les intervalles de confiance à 90% sont présentés pour chaque coefficient dans les deux modèles, illustrant la stabilité des estimations après correction.	44
3.15	Arbre de régression illustrant la relation entre les variables catégorielles et le salaire net médian. Chaque nœud de l'arbre représente une scission basée sur une variable, divisant les observations en groupes homogènes en termes de salaire.	47

Liste des tableaux

2.1	Statistiques descriptives par domaine académique	26
2.2	Statistiques descriptives par sexe académique	27
3.1	Régression linéaire pondérée	36

REMERCIEMENTS

Nous tenons à exprimer notre profonde gratitude à notre professeur référent, Christophe CHESNEAU, pour son soutien inestimable, ses conseils éclairés et son encouragement tout au long de la réalisation de ce projet. Son expertise et sa pédagogie ont été des atouts majeurs qui ont grandement contribué à l'aboutissement de notre recherche. Nous sommes également reconnaissants envers l'ensemble du corps professoral du Master 2 en Statistique pour leur enseignement rigoureux et leur disponibilité. Leurs connaissances et leur passion pour la statistique nous ont inspirés et guidés tout au long de notre parcours académique.

Enfin, nous souhaitons remercier nos camarades de classe pour l'esprit de collaboration et le soutien mutuel qui ont prévalu durant nos études. Cette expérience enrichissante nous a non seulement apporté des connaissances académiques, mais a également tissé des liens d'amitié et de respect.

Chapitre 1

Introduction

1.1 Contexte du projet

Dans un contexte mondialisé où l'éducation supérieure est à la fois un vecteur de mobilité sociale et un moteur de compétitivité économique, l'insertion professionnelle des diplômés de Master se positionne au cœur des préoccupations stratégiques des institutions éducatives, des étudiants, ainsi que des acteurs du marché du travail. L'envergure de ce défi est amplifiée par les mutations technologiques rapides et les évolutions sociétales qui transforment les besoins en compétences et redessinent les perspectives de carrière. La présente étude se propose d'analyser en profondeur les données relatives à l'insertion professionnelle des diplômés de Master en France, à travers le prisme d'indicateurs clés tels que le taux d'emploi et le salaire net médian, en mettant l'accent sur la variabilité interdisciplinaire de ces indicateurs et sur les disparités potentielles qu'ils révèlent.

1.2 Source des données

Les données analysées dans ce rapport proviennent des enquêtes nationales sur l'insertion professionnelle des diplômés de Master, organisées suite aux réformes législatives qui ont renforcé la mission des universités dans ce domaine. La loi relative aux libertés et responsabilités des universités (LRU) de 2007, suivie de la loi sur l'enseignement supérieur et la recherche (ESR) de 2013 et de la loi Orientation et Réussite des Étudiants (ORE) en 2018, ont toutes contribué à formaliser cette mission **Rose2014**. Ces lois ont été le catalyseur de la mise en place des enquêtes nationales par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, visant à évaluer l'insertion des diplômés 18 et 30 mois après leur diplomation.

Cette orientation vers une meilleure insertion professionnelle s'inscrit dans une dynamique plus large de professionnalisation des formations dans l'enseignement supérieur, marquée par l'introduction systématique de stages, ainsi

que par le développement de dispositifs d'orientation et d'accompagnement pour les étudiants [MÉNARD](#). Ces données, couvrent la période de 2011 à 2020, offrent un aperçu de l'évolution des conditions d'emploi et salariales des diplômés.

1.3 Objectifs de l'étude

Cette étude s'attache à décrypter les mécanismes d'insertion professionnelle des diplômés de Master dans le contexte français actuel, caractérisé par des transformations économiques et technologiques profondes. Elle vise à fournir une analyse détaillée des indicateurs d'emploi et de rémunération, explorant comment les facteurs éducatifs, démographiques et législatifs se conjuguent pour façonner les trajectoires professionnelles des jeunes diplômés.

Les objectifs spécifiques de l'étude sont les suivants :

- **Analyser les taux d'emploi et les salaires médians** : Quantifier et interpréter les taux d'emploi et les salaires médians pour différentes disciplines, en cherchant à identifier les variables clés qui influencent l'employabilité et le succès économique des diplômés.
- **Évaluer l'impact des politiques éducatives** : Examiner dans quelle mesure les réformes législatives et les politiques d'orientation et d'accompagnement ont contribué à l'amélioration de l'insertion professionnelle des diplômés de Master.
- **Informers les décisions stratégiques** : Fournir des données précieuses aux institutions éducatives, aux décideurs et aux étudiants pour guider les initiatives visant à renforcer l'adéquation entre formation universitaire et besoins du marché du travail.
- **Enrichir la recherche académique** : Offrir une contribution empirique et méthodologique à la littérature sur l'insertion professionnelle, en mettant en avant des analyses quantitatives et qualitatives poussées.
- **Servir de base à des études prospectives** : Poser les jalons pour des recherches futures, notamment des études longitudinales qui pourraient suivre l'évolution à long terme de l'insertion professionnelle des diplômés.

L'ambition de cette étude est de tracer un panorama exhaustif de l'insertion professionnelle des diplômés de Master, et de proposer des pistes de réflexion pour les défis actuels et futurs liés à l'éducation supérieure et au marché de l'emploi.

1.4 Méthodologie

Cette section détaille les méthodes statistiques et analytiques employées pour examiner les données relatives à l'insertion professionnelle des diplômés

de Master. Nous adoptons une approche méthodique pour assurer la rigueur et la précision de notre analyse.

1.4.1 Analyses descriptives

L'étape initiale de notre analyse se concentre sur l'examen approfondi de la distribution des diplômés, en se focalisant sur des aspects clés tels que l'année d'obtention du diplôme et la discipline d'étude. Cette phase descriptive joue un rôle fondamental dans notre compréhension globale du paysage de l'insertion professionnelle des diplômés de Master. Grâce à cette approche, nous dégagons des tendances générales et définissons la composition du groupe de diplômés, fournissant ainsi une base solide pour les investigations statistiques avancées.

Concrètement, nous mettons en évidence les variations des salaires nets médians, un indicateur du succès professionnel post-diplôme, révélant des différences de rémunération entre les divers domaines d'études. L'analyse des taux d'emploi offre un aperçu précieux sur les dynamiques du marché du travail et sur la capacité des parcours éducatifs à préparer efficacement les étudiants à l'emploi.

Un aspect de nos analyses descriptives est l'examen des disparités basées sur le genre, mettant en lumière les inégalités potentielles dans les trajectoires professionnelles et salariales des diplômés. Cette dimension de notre étude est essentielle pour comprendre les défis spécifiques rencontrés par les femmes et les hommes dans le marché du travail actuel.

Pour mener ces analyses, nous utilisons diverses méthodologies statistiques et outils de visualisation, permettant une exploration et une présentation claire des données. Les visualisations graphiques, comme les histogrammes et les diagrammes en boîte, jouent un rôle dans l'illustration des distributions et des tendances salariales parmi les diplômés.

1.4.2 Analyses bivariées et multivariées

Notre démarche analytique s'étend également aux analyses bivariées et multivariées, cruciales pour décrypter les relations complexes entre diverses variables et pour élucider les mécanismes sous-jacents régissant l'insertion professionnelle des diplômés. Ces analyses contribuent significativement à notre compréhension des interactions et des tendances au sein du marché du travail.

Les analyses bivariées, en se concentrant sur les relations entre deux variables à la fois, ont révélé des corrélations et des disparités importantes. En examinant, par exemple, les taux d'emploi et les salaires nets médians selon le genre, nous avons identifié des écarts marqués, mettant en lumière des inégalités de genre persistantes dans le monde professionnel. L'étude des écarts salariaux entre les différents domaines d'études a, de même, mis en avant l'influence prépondérante du choix de domaine sur les perspectives économiques des diplômés.

L'adoption d'analyses multivariées, notamment à travers l'Analyse en Composantes Principales (ACP), enrichit notre exploration en dévoilant les dynamiques entre plusieurs variables simultanément. L'ACP a permis de distinguer des axes principaux regroupant des variables associées à la qualité et à la stabilité de l'emploi, dessinant une cartographie intégrée des facteurs essentiels à l'insertion professionnelle. Cette approche multidimensionnelle offre une perspective approfondie sur les multiples facettes influençant l'employabilité et les niveaux de rémunération des diplômés, soulignant la complexité et l'interdépendance des facteurs à l'œuvre.

1.4.3 Modélisation statistique

Notre analyse s'appuie sur des techniques de modélisation statistique pour explorer l'impact de divers facteurs sur le salaire des diplômés. Nous utilisons la régression linéaire pour identifier les relations linéaires entre les variables et l'arbre de régression pour examiner les relations non linéaires et les interactions complexes entre les variables.

Arbre de régression

L'incorporation de l'arbre de régression dans notre étude vise à analyser comment des caractéristiques telles que le genre, le domaine d'études et la catégorisation basée sur la première composante principale de l'Analyse en Composantes Principales (ACP) influencent de manière non linéaire le salaire. Cette méthode révèle les dynamiques subtiles et les interactions entre variables qui ne seraient pas évidentes dans des modèles strictement linéaires.

Les avantages principaux de l'arbre de régression comprennent : - **Interprétabilité** : Ils fournissent un modèle visuellement accessible, où les décisions à chaque nœud sont explicitement définies, facilitant la compréhension des facteurs influençant le salaire net médian. - **Flexibilité** : Capables de modéliser efficacement des relations non linéaires et des interactions sans nécessiter la spécification préalable d'une forme fonctionnelle. - **Adaptabilité** : L'arbre s'ajuste aux données, identifiant les variables les plus influentes et révélant la structure sous-jacente affectant la variable à expliquée.

1.4.4 Tests statistiques

La rigueur de notre étude repose sur l'utilisation approfondie de tests statistiques pour examiner et valider les hypothèses sous-jacentes à nos modèles de régression. Ces tests assurent la fiabilité et la validité de nos conclusions. Parmi les tests appliqués, on compte :

- Le **test de Shapiro-Wilk** pour évaluer la normalité des résidus, une hypothèse fondamentale pour l'application de nombreux tests statistiques. Ce test

a aidé à confirmer que les résidus de nos modèles de régression suivent une distribution approximativement normale, validant ainsi les techniques d'inférence statistique utilisées. - Le **test Rainbow** pour vérifier la linéarité de la relation entre les variables indépendantes et la variable dépendante. Une hypothèse de linéarité valide est essentielle pour l'interprétation correcte des coefficients de régression. - Les tests d'**autocorrélation**, notamment le test de Durbin-Watson et le test de Ljung-Box, ont été utilisés pour détecter la présence d'autocorrélation dans les résidus, ce qui indique des erreurs dans la spécification du modèle. - Le **test de Breusch-Pagan** a servi à évaluer l'homogénéité de la variance des erreurs (homoscédasticité), une condition nécessaire pour que les estimations des erreurs standards soient fiables. - L'analyse de la **multicolinéarité** à travers le calcul des facteurs d'inflation de la variance (VIF) a permis de s'assurer que nos variables indépendantes ne sont pas excessivement corrélées, ce qui compromet l'interprétation des résultats.

1.4.5 Clustering et classification

Segmenter les diplômés selon leurs caractéristiques d'insertion professionnelle dévoile la complexité de leurs trajectoires. L'analyse de classification hiérarchique, en particulier, distingue des groupes aux profils d'insertion similaires, révélant ainsi les diverses dynamiques du marché du travail et les influences communes à chaque cluster.

Identifier ces groupes enrichit notre compréhension des mécanismes d'insertion et souligne l'impact de stratégies ciblées pour faciliter l'accès au marché du travail. Cette classification affine l'analyse, permettant d'élaborer des recommandations adaptées à chaque profil d'insertion et d'orienter efficacement les politiques d'accompagnement vers le succès professionnel.

Chapitre 2

Analyse descriptive

2.1 Au coeur des disparités salariales : Ce que les salaires médians nous révèlent

Les salaires médians par domaine d'études sont des indicateurs cruciaux de l'équité et de l'efficacité du marché de l'emploi. En scrutant de près ces données, nous pouvons saisir l'essence des disparités salariales qui persistent dans notre société. Cette section explore en profondeur les nuances des salaires médians, révélant des tendances fascinantes et parfois inquiétantes qui émergent lorsque l'on examine de près les écarts entre les différents secteurs et entre les sexes.

À travers une analyse détaillée des données sur les salaires médians par domaine et par sexe, nous mettrons en lumière les dynamiques complexes qui sous-tendent ces disparités salariales. En examinant les variations de rémunération entre les différents secteurs d'études, nous chercherons à comprendre les facteurs qui influent sur les perspectives économiques des diplômés et des professionnels. De même, en explorant les différences salariales entre hommes et femmes au sein de chaque domaine, nous aborderons les questions cruciales de l'égalité des sexes et de la discrimination salariale.

2.1.1 Quels facteurs sous-tendent les différences de salaires médians par domaine d'études ?

La répartition des salaires médians par domaine est un indicateur clé de l'insertion professionnelle des diplômés. Elle offre une vue d'ensemble des perspectives économiques associées à chaque domaine d'études.

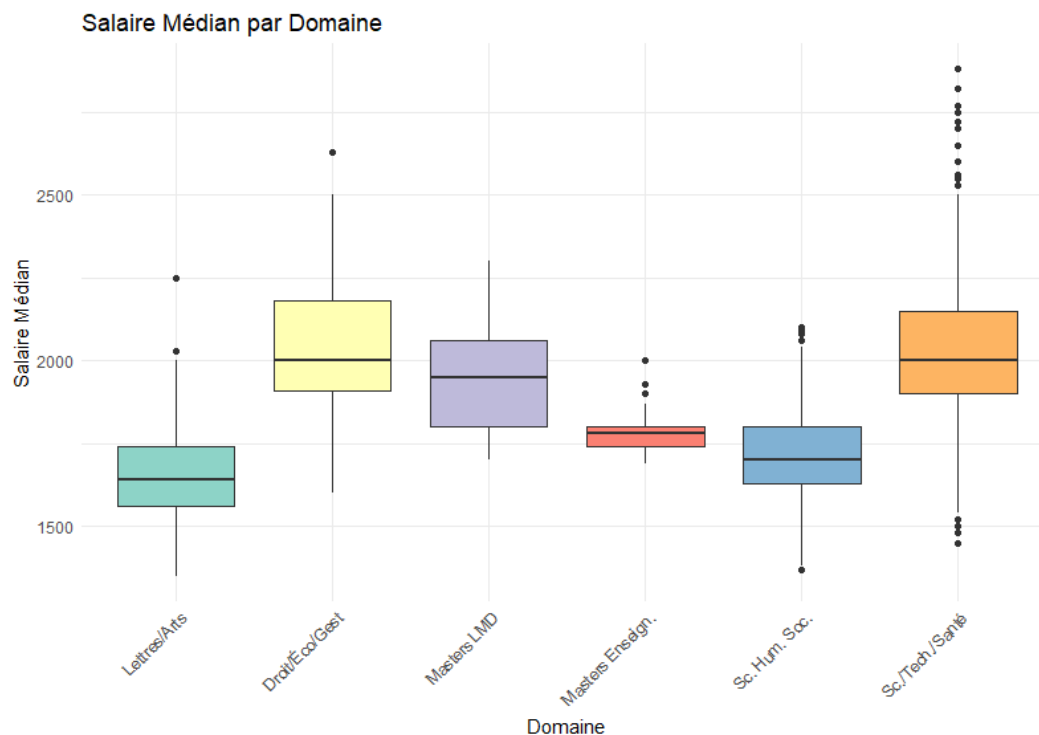


FIGURE 2.1 – Salaire Médian par Domaine

La Figure 2.1 présente les salaires médians des diplômés par domaine d'étude. Il est notable que les domaines liés à la Santé et à la Science/Technologie/Santé se distinguent par des médianes salariales supérieures, ce qui reflète la demande du marché et l'investissement en compétences spécialisées dans ces secteurs. Les domaines des Lettres/Arts et des Sciences Humaines affichent des médianes inférieures, ce qui est souvent observé dans les tendances économiques actuelles. L'écart interquartile large dans des domaines comme la Science/Technologie/Santé suggère une hétérogénéité significative dans les salaires, possiblement attribuable à une grande diversité de carrières au sein de ces domaines. Les observations qui se situent en dehors des moustaches du boxplot indiquent des salaires exceptionnellement élevés, qui signifient, bien que moins communs, il existe des postes très rémunérateurs dans presque tous les domaines étudiés et cela est très aigu dans le domaine de la Science/Technologie/Santé.

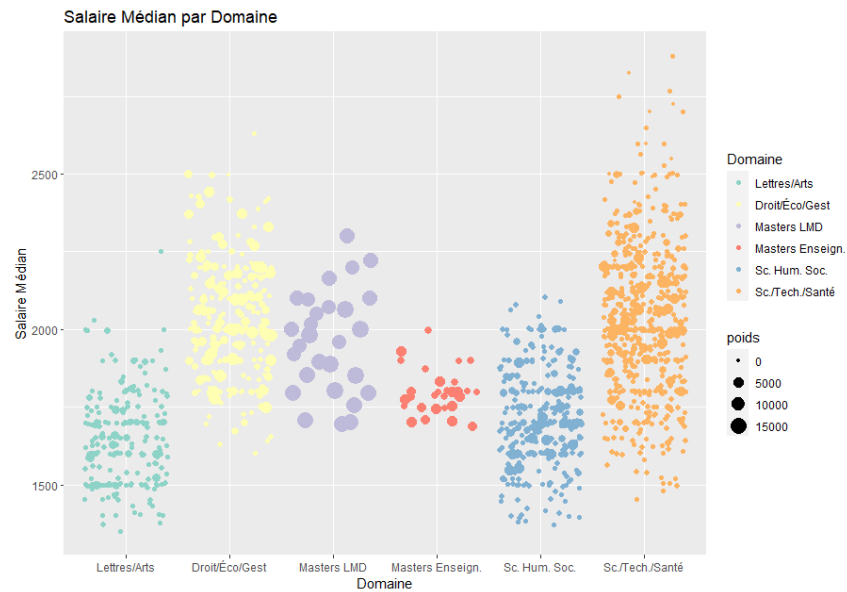


FIGURE 2.2 – Détail du Salaire Médian par Domaine

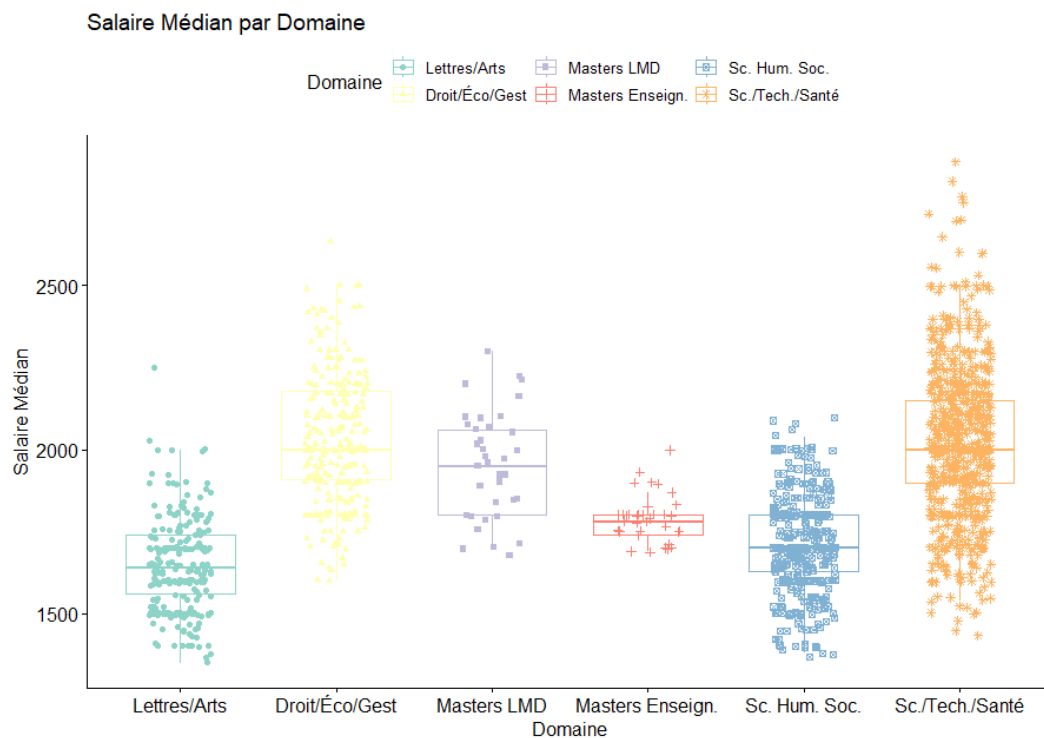


FIGURE 2.3 – Détail du Salaire Médian par Domaine

La Figure 2.3 offre une perspective plus détaillée sur la distribution des salaires médians par domaine, en tenant compte de la pondération des réponses.

Cette représentation en nuage de points permet d'apprécier non seulement les médianes salariales, mais aussi la concentration des réponses et leur poids relatif, ce dernier étant indiqué par la taille des points et représentant le nombre de réponses par secteur disciplinaire.

On observe une concentration plus dense de points dans les domaines des Lettres/Arts et des Sciences Humaines et Sociales, suggérant un volume de données substantiel et une estimation potentiellement plus robuste de la médiane salariale pour ces domaines. Ces domaines semblent donc non seulement présenter des salaires médians inférieurs mais aussi une plus grande homogénéité en termes de réponses collectées.

Les domaines du Droit/Économie/Gestion et des Sciences/Technologie/Santé, en revanche, montrent non seulement une variabilité plus importante des salaires médians mais également des points de pondération plus élevés, indiquant une prévalence de réponses plus conséquente. Cela pourrait suggérer que les salaires dans ces domaines sont influencés par un éventail plus large de facteurs ou que ces domaines attirent un nombre plus important de répondants à haut salaire.

La variété dans la taille des points à travers le graphique illustre la diversité des situations professionnelles au sein de chaque domaine. Les points plus importants reflètent un plus grand nombre de réponses et donc une influence plus marquée sur la médiane salariale calculée. Il convient de noter que les domaines tels que les Sciences/Technologie/Santé, malgré des salaires médians globalement plus élevés, révèlent une complexité et une variabilité sous-jacentes qui pourraient masquer des disparités significatives au sein de ces catégories.

Les inégalités salariales observées entre les différents domaines d'études peuvent être partiellement expliquées par les disparités des coûts de formation. Les chiffres publiés par la Direction de l'Évaluation, de la Prospective et de la Performance (DEPP) en 2021 révèlent que la dépense moyenne par étudiant en France était de 11 530 euros en 2019. Toutefois, cette moyenne masque des disparités notables : 10 110 euros sont dépensés en moyenne pour un étudiant à l'université, tandis que ce montant s'élève à 14 270 euros pour les étudiants en sections de techniciens supérieurs (STS) dans le domaine des Sciences Technologies et Santé, et atteint 15 710 euros pour ceux en classes préparatoires aux grandes écoles (CPGE) [DIRECTION DE L'ÉVALUATION, DE LA PROSPECTIVE ET DE LA PERFORMANCE \(DEPP\)](#).

Ces coûts reflètent les ressources investies dans la formation et sont susceptibles d'influencer les attentes salariales des diplômés. Les formations les plus onéreuses, souvent associées à des taux d'encadrement plus élevés et des infrastructures plus coûteuses, préparent généralement les étudiants à des emplois hautement qualifiés qui tendent à offrir de meilleurs salaires. À l'inverse, les domaines présentant des coûts de formation inférieurs peuvent conduire à des salaires de départ moins élevés, reflétant peut-être une saturation du marché ou des parcours professionnels moins directement liés à des postes spécialisés.

En plus des coûts de formation proprement dits, les différences dans les salaires médians par domaine peuvent également être affectées par la sélectivité des programmes, le contenu et les méthodes d'apprentissage, ainsi que par l'évolution de la demande pour certaines compétences sur le marché du travail [FACK et HUILLERY](#).

2.1.2 Quels mystères se dissimulent derrière les inégalités de salaires entre les sexes ?

Comme illustré dans la Figure [2.4](#), les données suggèrent que le salaire net médian des hommes était supérieur à celui des femmes chaque année consécutive. En 2013, les femmes gagnaient en moyenne 1735 euros par mois tandis que les hommes gagnaient 1973 euros, établissant un écart initial de 238 euros. Cet écart s'est maintenu avec une légère augmentation dans les années suivantes, culminant à une différence de 219 euros en 2020, où les femmes gagnaient 1958 euros contre 2177 euros pour les hommes. Les études historiques et récentes montrent que cet écart résulte d'une combinaison complexe de facteurs structurels et individuels.

D'après Meurs et Ponthieux (2000), l'écart salarial global en 1997 était de 27% en faveur des hommes, une différence attribuable en grande partie à la durée hebdomadaire de travail et à d'autres facteurs structurels. Ils estiment que sur cet écart, deux cinquièmes sont expliqués par les différences de durée de travail, principalement due au temps partiel plus fréquent chez les femmes, et deux autres cinquièmes par d'autres différences structurelles, laissant un cinquième inexpliqué, qui pourrait suggérer l'existence d'une discrimination salariale [MEURS et PONTHEUX](#).

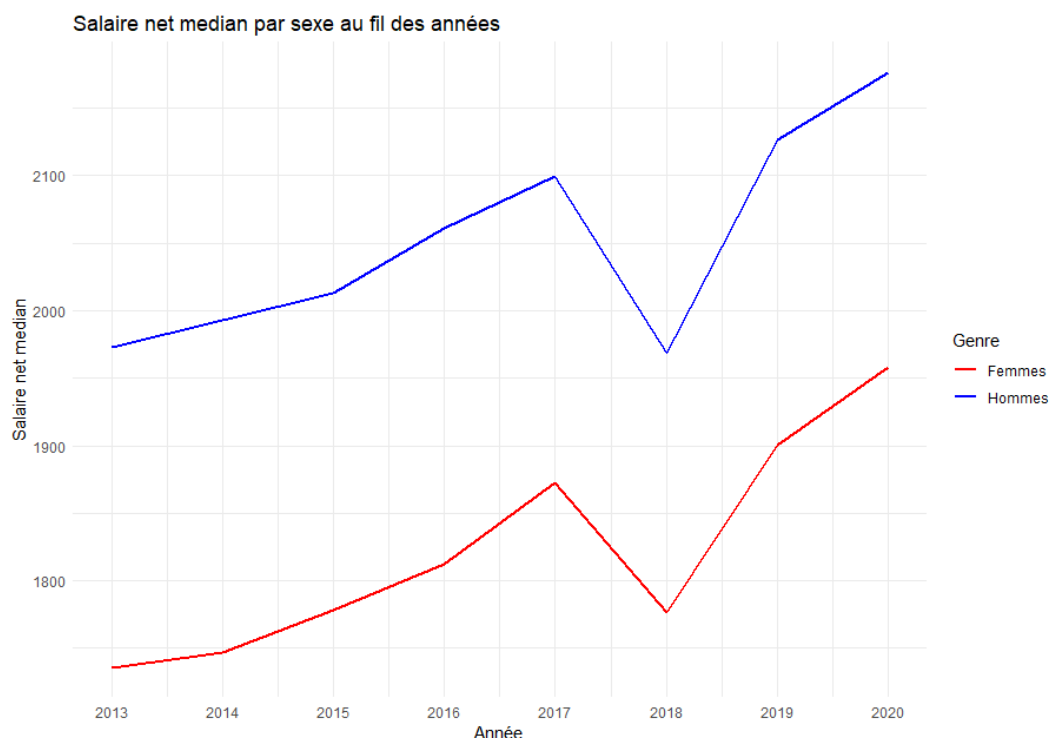


FIGURE 2.4 – Évolution de l'écart de salaire net moyen entre les sexes de 2013 à 2020

En 2020, bien que l'écart salarial pour les salariés à temps complet se soit réduit à 11%, la part inexpliquée, potentiellement attribuable à la discrimination, reste significative. La Figure 2.4 illustre ces différences continues de rémunération.

Les recherches de Georges-Kot (2020) indiquent que les inégalités de salaire pour un même volume de travail ont diminué d'un quart au cours des vingt dernières années, grâce en partie à une réduction des écarts de volume de travail. Cependant, les inégalités persistent, notamment en raison de la concentration des femmes dans certaines catégories d'emploi et de leur sous-représentation dans les postes les mieux rémunérés, en particulier parmi les salariés ayant des enfants [GEORGES-KOT](#).

La persistance de ces inégalités souligne l'importance d'aborder les différences de secteur et d'emploi occupé, ainsi que la nécessité de politiques ciblées pour promouvoir l'égalité des chances et traiter les causes profondes de l'écart de rémunération entre les sexes.

2.1.3 Quels domaines d'études exacerbent le plus les inégalités salariales entre hommes et femmes ?

La Figure 2.5 met en évidence l'écart persistant de salaire net médian entre hommes et femmes, révélant des disparités particulièrement marquées dans certains domaines d'études. Notamment, les secteurs où cet écart est le plus prononcé, tels que Droit/Économie/Gestion et Sciences/Technologie/Santé, correspondent également à ceux caractérisés par un taux élevé d'emplois cadres et une prédominance des débouchés vers le secteur privé.

Cette corrélation peut être expliquée par plusieurs facteurs. Tout d'abord, les postes de cadre, souvent plus nombreux dans le secteur privé, tendent à offrir des salaires plus élevés en raison des compétences spécialisées requises et de la responsabilité accrue associée à ces rôles. Cependant, ces postes sont historiquement et structurellement plus accessibles aux hommes, contribuant ainsi à l'écart salarial observé. De plus, le secteur privé, connu pour sa compétitivité salariale, peut aggraver ces écarts par des politiques de rémunération plus agressives qui favorisent les profils traditionnellement masculins.

Par ailleurs, ces domaines d'études sont souvent caractérisés par des négociations salariales et des progressions de carrière plus dynamiques, ce qui peut entraîner une différenciation accrue des salaires. Malgré des qualifications équivalentes, les femmes se heurtent souvent à des obstacles structurels qui limitent leur accès aux postes les mieux rémunérés, tels que le plafond de verre et les biais inconscients lors des processus de recrutement et de promotion.

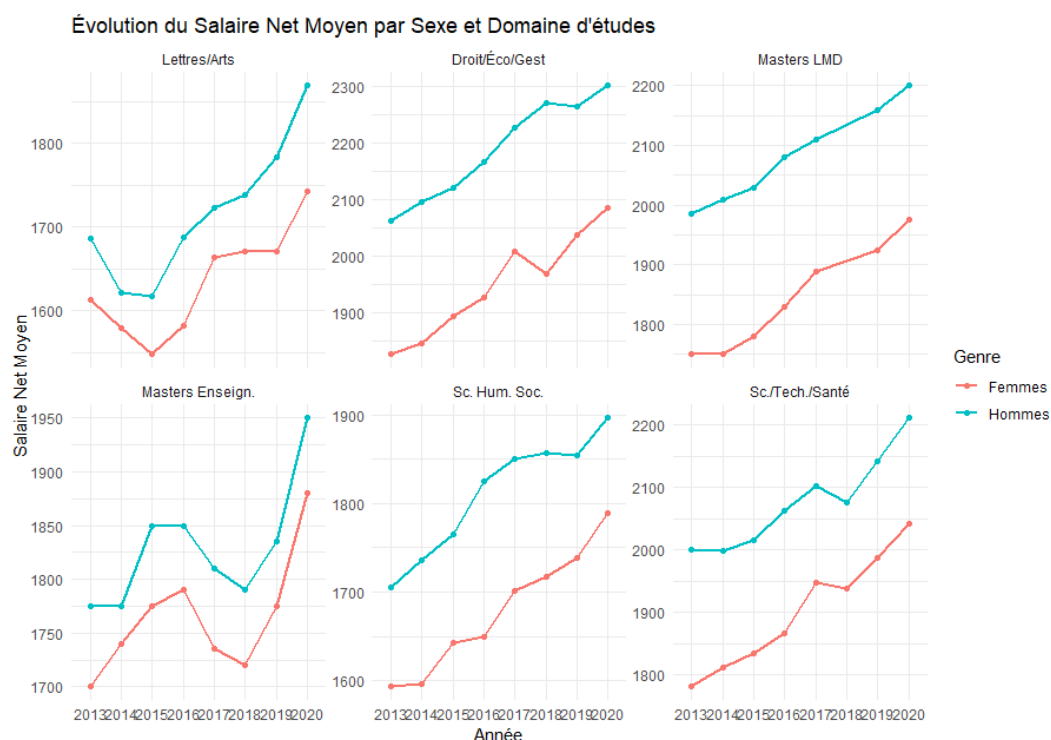


FIGURE 2.5 – Évolution de l'écart de salaire médian net entre les sexes de 2013 à 2020

Il est à noter que ces inégalités salariales sont moins prononcées dans la fonction publique que dans le secteur privé. En 2017, l'écart de salaire en Équivalent Temps Plein (EQTP) entre hommes et femmes était de 12,4 % dans la fonction publique, contre 16,8 % dans le secteur privé [GEORGES-KOT](#).

La ségrégation professionnelle joue également un rôle significatif dans l'écart de salaire entre les sexes. Les femmes se concentrent souvent dans un nombre restreint de professions, tandis qu'une proportion plus faible d'entre elles occupent des postes de cadre. Cette tendance est exacerbée par des orientations scolaires différenciées selon le sexe, ce qui a des répercussions durables sur les inégalités salariales [GEORGES-KOT](#).

2.1.4 Les différences de salaires médians par domaine : sont-elles statistiquement significatives ?

Avant d'examiner les différences de salaires médians par domaine, il est judicieux de déterminer si ces différences sont statistiquement significatives. À cette fin, le test non paramétrique de Kruskal-Wallis est utilisé. Ce test permet de comparer les médianes de deux groupes ou plus pour déterminer s'il existe des différences significatives entre eux. Contrairement à l'ANOVA, le test de Kruskal-Wallis ne suppose pas une distribution normale des résidus

Salaire Médian par Domaine

Domaine

- Lettres/Arts
- Droit/Éco/Gest
- Masters LMD
- Masters Enseign.
- Sc. Hum. Soc.
- Sc./Tech./Santé

Salaire Médian

Kruskal-Wallis, $p < 2.2e-16$

Significance levels for pairwise comparisons:

- Lettres/Arts vs Droit/Éco/Gest: $p < 2.22e-16$
- Lettres/Arts vs Masters LMD: $p < 2.22e-16$
- Lettres/Arts vs Masters Enseign.: $p < 2.22e-16$
- Lettres/Arts vs Sc. Hum. Soc.: $p < 2.22e-16$
- Lettres/Arts vs Sc./Tech./Santé: $p < 2.22e-16$
- Droit/Éco/Gest vs Masters LMD: $1.8e-07$
- Droit/Éco/Gest vs Masters Enseign.: $5.5e-06$
- Droit/Éco/Gest vs Sc. Hum. Soc.: 0.54
- Droit/Éco/Gest vs Sc./Tech./Santé: $4.7e-11$
- Masters LMD vs Masters Enseign.: $5.5e-06$
- Masters LMD vs Sc. Hum. Soc.: 0.54
- Masters LMD vs Sc./Tech./Santé: $4.7e-11$
- Masters Enseign. vs Sc. Hum. Soc.: 0.54
- Masters Enseign. vs Sc./Tech./Santé: $4.7e-11$
- Sc. Hum. Soc. vs Sc./Tech./Santé: $p < 2.22e-16$

La Figure 2.6 illustre les résultats de l'application du test de Kruskal-Wallis aux salaires médians par domaine d'étude. Comme le révèlent les annotations sur le graphique, les différences de salaires médians entre certains domaines sont statistiquement significatives, avec des valeurs de p inférieures au seuil conventionnel de 0.05. Ces résultats indiquent une variabilité dans le potentiel de revenu en fonction du domaine d'études choisi, avec une distinction notable entre des domaines comme les Lettres/Arts et les Sciences/Technologie/Santé, ce dernier affichant des salaires médians nettement plus élevés.

Page 15

2.2 Sur les chemins de l'emploi : une analyse du taux d'insertion professionnelle

L'analyse du taux d'insertion professionnelle offre un éclairage précieux sur la transition des diplômés vers le marché du travail. En examinant les tendances et les facteurs qui influencent cette transition, nous pouvons mieux comprendre les défis et les opportunités auxquels sont confrontés les nouveaux diplômés. Cette section explore les différents chemins empruntés par les diplômés universitaires et les implications de ces parcours pour leur intégration professionnelle.

”Quelle voie vers l'emploi ? : Analyse du taux d'insertion professionnelle”

2.2.1 Évolution des taux d'insertion professionnelle par sexe : Analyse des tendances et des disparités sur de 8 ans”

L'analyse des taux d'insertion par genre au fil des années révèle des tendances significatives et des écarts entre les sexes. Ces écarts peuvent éclairer les discussions sur l'égalité des genres dans l'accès à l'emploi après l'obtention d'un diplôme.

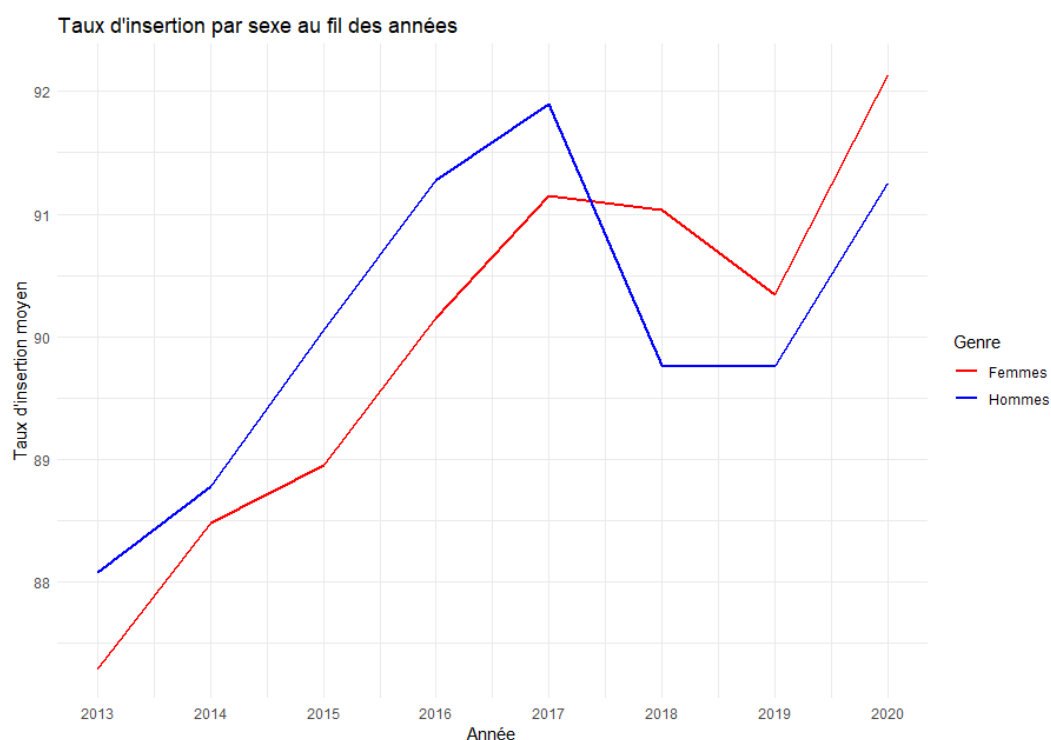


FIGURE 2.7 – Taux d'insertion par sexe au fil des années

Comme illustré dans la Figure 2.7, un écart persistant entre les femmes

et les hommes est visible sur la période de 2013 à 2020. En 2013, le taux d'insertion moyen des femmes était de 87.3%, tandis que celui des hommes s'élevait à 88.1%, marquant un écart de 0.8 points de pourcentage. Cet écart s'est légèrement resserré en 2014 avec un taux d'insertion des femmes à 88.5% contre 88.8% pour les hommes.

Toutefois, l'écart s'est accentué en 2015, atteignant 1.1 points de pourcentage, avec un taux d'insertion des femmes à 89.0% comparé à 90.1% pour les hommes. Les années suivantes ont montré des variations, avec un écart notable en 2017 de 0.8 points de pourcentage, avant une inversion surprenante en 2018 où les femmes ont dépassé les hommes avec un taux d'insertion de 91.0% contre 89.8%.

Comme l'ont souligné Couppié et Epiphane COUPPIÉ et EPIPHANE, [“Et les femmes devinrent plus diplômées que les hommes...”](#) au cours des deux dernières décennies, la situation des jeunes femmes sur le marché du travail s'est améliorée de manière significative : elles sont plus diplômées, plus présentes en emploi et bénéficient d'un début de rattrapage salarial. Elles accèdent également à des métiers et des filières plus proches de ceux des hommes, bien que des inégalités persistent, notamment dans l'accès au statut de cadre.

Par ailleurs, la féminisation des emplois occupés par les jeunes, un phénomène observé depuis plusieurs décennies sur le marché du travail français, a été soutenue par la tertiarisation et la montée du niveau de qualification des emplois en France [“La relation genre-insertion at-elle évolué en 20 ans ?”](#) Cette tendance se reflète également dans la réduction de la ségrégation professionnelle, avec une baisse notable de la part des emplois très féminisés ou très masculinisés, au profit des métiers de composition plus mixte. Ces observations sont cohérentes avec les tendances que nous avons identifiées dans notre analyse, soulignant l'importance de considérer les dynamiques de genre dans les politiques d'insertion professionnelle.

2.2.2 Plongée dans l'insertion professionnelle par domaine : révélations sur les tendances évolutives et les disparités cachées

Les fluctuations des taux d'insertion par domaine académique au fil des années peuvent refléter une variété de facteurs économiques et politiques éducatifs. Ces tendances sont essentielles pour évaluer l'adéquation des formations avec les besoins du marché du travail.

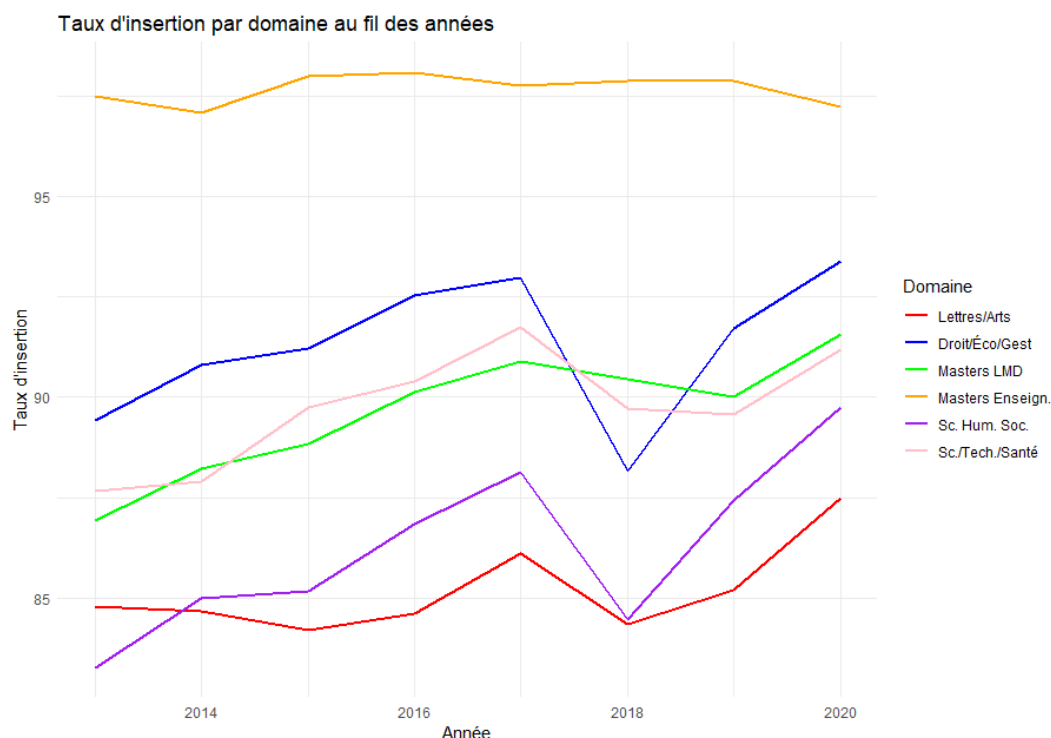


FIGURE 2.8 – Taux d'insertion par domaine au fil des années

Comme le montre la Figure 2.8, il existe des écarts significatifs entre les domaines, ainsi que des variations notables au sein de chaque domaine au fil des ans. Par exemple, le domaine des Sciences Humaines et Sociales a montré une augmentation constante, passant de 83.25% en 2013 à 89.75% en 2020, ce qui pourrait indiquer une meilleure adaptation de ces diplômés au marché du travail ou une amélioration de la qualité de l'enseignement dans ce domaine.

En revanche, le domaine de Droit/Économie/Gestion, bien qu'il ait commencé avec un taux d'insertion relativement élevé en 2013 de 89.43%, a connu une baisse en 2018 à 88.15%, suivie d'une reprise à 93.41% en 2020. Cette baisse temporaire pourrait être due à des facteurs externes tels que les changements réglementaires ou les fluctuations économiques.

Les Lettres et les Arts ont également connu des fluctuations, avec une baisse en 2018 à 84.33%, mais une reprise remarquable à 87.47% en 2020. Ces mouvements peuvent refléter les changements culturels et les initiatives visant à valoriser les compétences dans ces domaines.

Sciences/Technologie/Santé montre une augmentation des taux d'insertion à partir de 2019. Cette tendance à la hausse peut être attribuée à la montée en puissance des métiers liés à l'informatique, tels que le développement de logiciels, le big data, et le machine learning. L'accent mis sur la numérisation et l'innovation technologique a entraîné une demande accrue pour les compétences associées à ces secteurs. De plus, l'urgence de problématiques sanitaires mon-

diales a probablement stimulé la nécessité de compétences avancées en santé et en analyse de données.

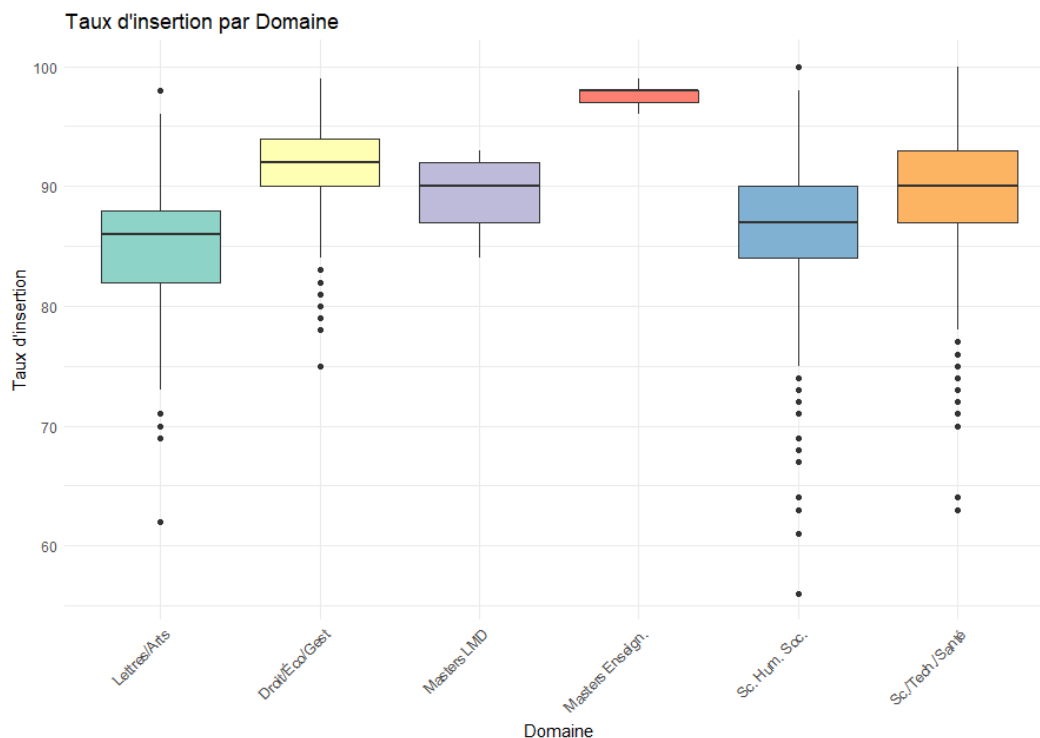


FIGURE 2.9 – Distribution des taux d'insertion par domaine

Dans la Figure 2.9, le domaine des Lettres/Arts montre une médiane inférieure et une variabilité plus importante des taux d'insertion par rapport aux autres domaines. La présence de valeurs aberrantes significatives en dessous du premier quartile indique que certaines sections disciplinaires au sein des Lettres/Arts et en Sciences humaines et sociales ont des taux d'insertion exceptionnellement bas. Ces valeurs reflètent des sous-disciplines telles que l'archéologie, l'ethnologie et la préhistoire où le passage à un emploi stable est moins direct, peut-être en raison de chemins de carrière atypiques, d'une compétition accrue pour des postes limités, ou d'un manque d'adéquation entre les compétences enseignées et celles demandées par le marché du travail.

Le domaine de Droit/Éco/Gest, quant à lui, bien qu'il présente des taux d'insertion relativement élevés, dénote également une certaine variabilité, ce qui montre une différence dans l'employabilité selon les spécialisations au sein du domaine. Les Masters LMD montrent une distribution plus concentrée, indiquant une uniformité relative dans l'insertion professionnelle.

Les Sciences/Technologie/Santé, en particulier, affichent une tendance à la hausse des taux d'insertion, ce qui est attribué à l'importance croissante de l'expertise technique dans l'économie moderne. La forte demande pour les

compétences en informatique, en analyse de données et en biotechnologie, stimulée par la révolution numérique et les défis de santé globaux, explique ces taux élevés et la tendance positive observée.

Les données sur l'insertion professionnelle des diplômés révèlent des nuances importantes lorsqu'elles sont décomposées par domaine disciplinaire. Il apparaît que pour un diplôme identique, les taux d'insertion varient légèrement selon la discipline étudiée. Selon les données récentes, les diplômés de Master dans les domaines de Droit-Économie-Gestion (DEG) et de Sciences-Technologies-Santé (STS) affichent un taux d'emploi supérieur à 90 % à 30 mois après l'obtention de leur diplôme. En comparaison, ce taux est légèrement inférieur pour leurs homologues en Sciences Humaines et Sociales (SHS) et en Lettres-Langues-Arts (LLA), s'établissant respectivement à 86 % et 87 % [MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION \(MENESR-DGESIP/DGRI-SIES\)](#)

2.3 Les emplois stables : Une réalité ou un mirage ?

Dans un paysage professionnel en constante évolution, la notion d'emploi stable reste un sujet de débat et d'interrogation. Alors que certains considèrent les emplois stables comme un objectif professionnel à atteindre pour assurer la sécurité financière et professionnelle, d'autres remettent en question la viabilité de cette perspective dans un monde caractérisé par la flexibilité et la précarité de l'emploi. Dans cette section, nous examinons les tendances du taux d'emplois stables au fil du temps, explorant les réalités changeantes du marché du travail et les implications pour les travailleurs et les employeurs.

2.3.1 Domaines offrant les emplois les plus stables : analyse des tendances récentes

L'analyse des taux d'emplois stables offre une perspective sur la sécurité de l'emploi dans divers domaines académiques après l'obtention d'un diplôme. Ces taux sont des indicateurs de la capacité des domaines à offrir des postes non seulement disponibles mais aussi pérennes.

La Figure 2.10 dévoile les tendances des taux d'emplois stables de 2013 à 2020. Le domaine des 'Masters Enseignement' montre un haut niveau de stabilité d'emploi, bien que le graphique indique une tendance à la baisse à partir de 2016.

Concernant les 'Sciences/Technologie/Santé', une hausse significative des taux d'emplois stables est observée depuis 2019. Cette tendance peut être interprétée comme le reflet de la forte demande pour les compétences en STEM

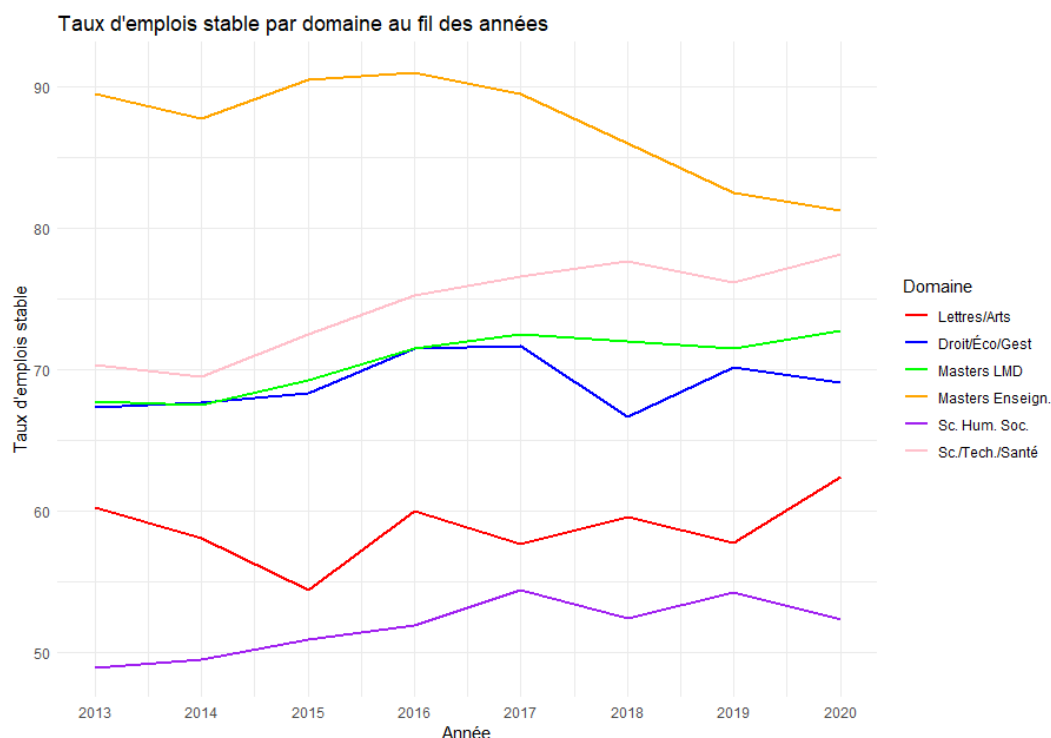


FIGURE 2.10 – Taux d’emplois stables par domaine au fil des années

(Science, Technologie, Ingénierie et Mathématiques), renforcée par la croissance des secteurs comme l’informatique, le big data et le machine learning.

Les 'Lettres/Arts', avec une certaine variabilité et des valeurs aberrantes, suggèrent une hétérogénéité dans la stabilité d’emploi au sein de ce domaine, potentiellement due à la nature des emplois dans le secteur culturel, souvent caractérisés par des contrats à court terme ou des projets ponctuels.

Les domaines de 'Droit/Éco/Gest' et 'Masters LMD' présentent des variations moins prononcées, ce qui reflète une relative stabilité dans ces secteurs professionnels, tandis que les 'Sciences Humaines et Sociales' montrent une augmentation progressive de la stabilité d’emploi, peut-être en raison de l’évolution des besoins dans le secteur des services et de la gestion des ressources humaines.

La différence de conditions d’emploi entre ces domaines est également notable. Les diplômés en 'Droit/Éco/Gest' et 'Sciences/Technologie/Santé' tendent à bénéficier d’emplois plus stables et à temps plein, avec des salaires nets médians mensuels supérieurs de 20 % par rapport à ceux obtenus par les diplômés en SHS et 'Lettres/Arts'. Cette tendance est soutenue par le fait que plus de neuf diplômés de STS sur dix en emploi occupent des postes de cadres ou de professions intermédiaires, un pourcentage significativement plus élevé que les 74 % à 84 % observés pour les autres domaines disciplinaires. Une explication partielle de ces écarts salariaux réside dans la plus grande

proportion d'emplois au sein du secteur public ou associatif pour les diplômés en Sciences Humaines et Sociales et 'Lettres/Arts', qui sont traditionnellement moins rémunérateurs que les postes dans le secteur privé [MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION \(MENESR-DGESIP/DGRI-SIES\)](#).

2.3.2 La stabilité de l'emploi : une disparité persistante entre les sexes

Une analyse du taux d'emplois stables par sexe au fil des années démontre des disparités notables entre les femmes et les hommes dans la stabilité de l'emploi qu'ils parviennent à sécuriser après l'obtention de leur diplôme.

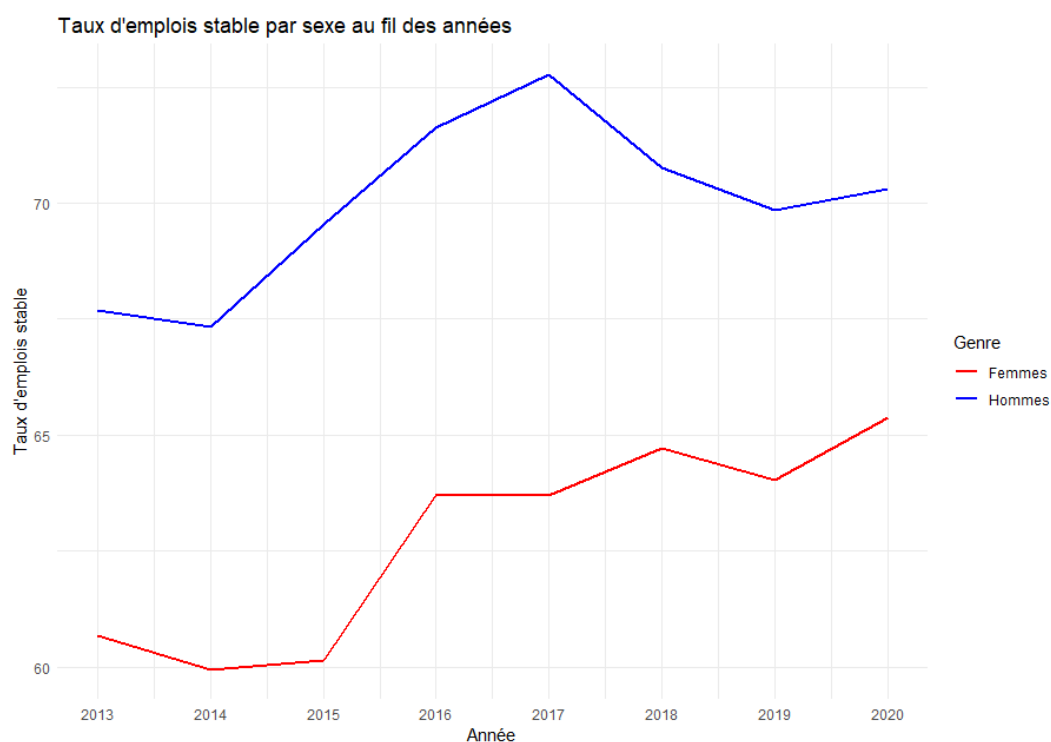


FIGURE 2.11 – Taux d'emplois stables par sexe au fil des années

La Figure 2.11 illustre un écart persistant entre les sexes en matière de sécurité de l'emploi, avec les hommes jouissant d'un taux supérieur d'emplois stables tout au long de la période étudiée. En 2013, les femmes avaient un taux d'emplois stables de 60.7%, tandis que les hommes avaient un taux de 67.7%. Cet écart s'est maintenu et a même légèrement augmenté, avec les femmes atteignant 65.4% et les hommes 70.3% en 2020.

Ces différences peuvent s'expliquer par une variété de facteurs structurels et sociétaux. Les discriminations de genre sur le lieu de travail, les inégalités

dans les opportunités de progression de carrière, et les différences dans les choix de domaine d'étude pourraient contribuer à cet écart. De plus, les femmes sont souvent plus affectées par les interruptions de carrière liées à la maternité et aux responsabilités familiales, ce qui peut affecter la continuité et la stabilité de leur emploi.

Les tendances montrent cependant une légère amélioration pour les femmes en 2016, où le taux d'emplois stables augmente de manière plus marquée, passant de 60.1% en 2015 à 63.7% en 2016. Cette progression, bien que positive, nécessite une action continue pour aborder les causes profondes des inégalités de genre et pour renforcer la stabilité de l'emploi pour toutes et tous.

Il est impératif pour les décideurs et les institutions d'enseignement de reconnaître ces disparités et de mettre en œuvre des politiques qui encouragent une plus grande équité en matière d'emploi. Des mesures telles que des programmes de mentorat, des initiatives de soutien à la maternité et à la parentalité, et une sensibilisation accrue à la diversité et à l'inclusion peuvent aider à réduire cet écart et à garantir que tous les diplômés aient accès à des emplois stables et rémunérateurs.

2.4 Analyse bivariée des disparités de sexe

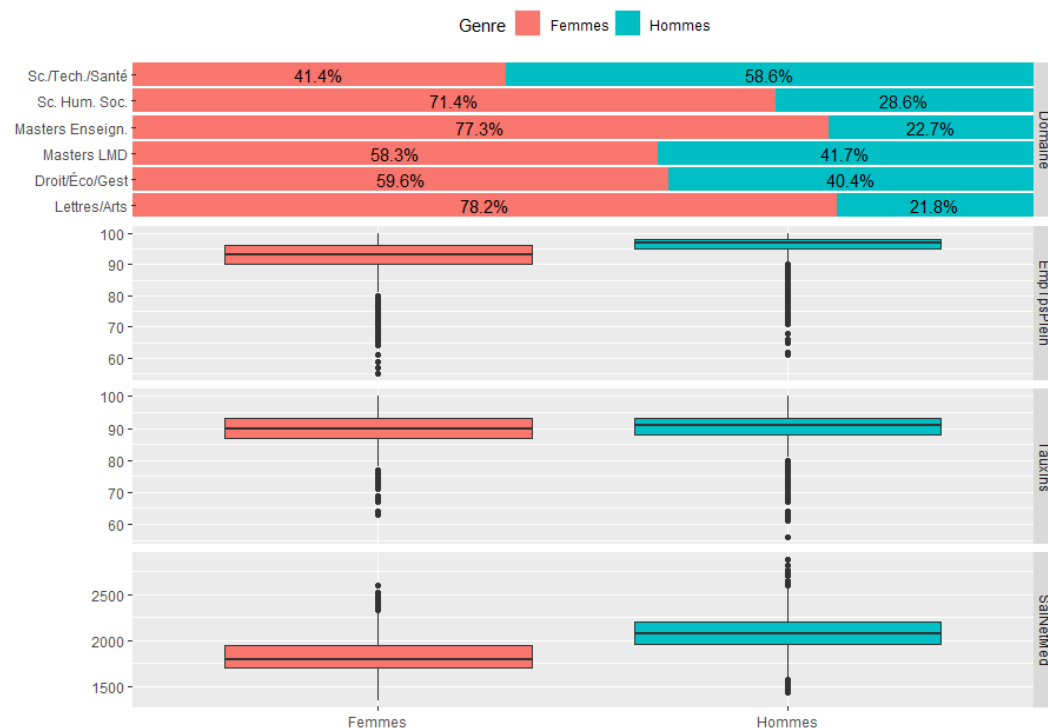


FIGURE 2.12 – Répartition des genres par domaine et distribution des emplois à temps plein et des salaires nets médians

La Figure 2.12 illustre de manière frappante les disparités de genre dans plusieurs domaines. Notamment, le domaine des Sciences Humaines et Sociales affiche une majorité écrasante de femmes, tandis que les hommes prédominent dans le domaine des Sciences/Technologie/Santé. Ces répartitions sont le reflet des orientations traditionnelles de genre, mais elles pourraient également indiquer où des efforts pourraient être faits pour promouvoir la diversité et l'équilibre des genres dans tous les domaines académiques.

Les distributions des emplois à temps plein et des salaires nets médians, indiquées par les boîtes à moustaches, révèlent des tendances et des écarts significatifs entre les sexes. Les femmes ont une médiane de salaire net inférieure à celle des hommes dans la plupart des domaines, une constatation qui soulève des questions sur l'équité salariale et l'égalité des chances professionnelles.

Les emplois à temps plein ne sont pas uniformément répartis entre les femmes et les hommes, ce qui peut influencer l'équilibre entre vie professionnelle et vie privée et, par extension, les choix de carrière à long terme.

2.5 Corrélations entre taux d'insertion, type d'Emploi, et sexe

Notre analyse statistique révèle des relations significatives entre le taux d'insertion, le type d'emploi, et le genre, comme le montre la matrice de corrélations ci-dessous.

Les coefficients de corrélation mesurent la force et la direction de la relation linéaire entre deux variables. Ils varient entre -1 et +1, où +1 indique une corrélation positive parfaite, -1 une corrélation négative parfaite, et 0 l'absence de corrélation. En général, les coefficients peuvent être interprétés comme suit :

- Forte corrélation : $|r| \geq 0.7$
- Corrélation modérée : $0.3 \leq |r| < 0.7$
- Faible corrélation : $|r| < 0.3$



FIGURE 2.13 – Matrice de corrélations entre taux d'insertion, type d'emploi, et genre

La Figure 2.13 affiche les coefficients de corrélation entre diverses variables liées à l'emploi et le genre. Par exemple, la corrélation entre le taux d'insertion et les emplois stables est de 0.686, indiquant une corrélation modérée à forte, particulièrement chez les hommes avec un coefficient de 0.724. Cela suggère que pour les hommes, une plus grande proportion d'emplois stables est associée à un taux d'insertion plus élevé.

Pour les femmes, bien que la corrélation entre le taux d'insertion et les emplois stables soit également positive (0.690), elle est légèrement inférieure à celle des hommes, ce qui peut indiquer des différences dans les facteurs qui contribuent à la stabilité de l'emploi entre les sexes.

Les nuages de points et les histogrammes en dessous des coefficients fournissent une représentation visuelle de la distribution et de la relation entre ces variables. Ils montrent une distribution plus large des taux d'insertion et des emplois stables parmi les hommes, ainsi qu'une tendance à des salaires nets médians plus élevés par rapport aux femmes.

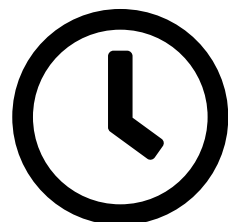
Résumé des statistiques



90.02% taux d'insertion professionnelle



63.6% emploi de niveau cadre



70.2% emploi à durée indéterminée



1928 salaire net médian

Résumé par domaine pour toutes ces années





Domaine	Statistiques			
	 Taux d'insertion	 Emploi cadre	 Emploi stable	 Salaire net
Lettres/Arts	85.2%	43.9%	60.8%	1649€
Droit/Éco/Gest	91.8%	60.8%	75.1%	2044€
Masters LMD	89.6%	62.3%	69.6%	1945€
Masters Enseign.	97.7%	86.1%	87.3%	1783€
Sc. Hum. Soc.	86.7%	57.7%	53.7%	1721€
Sc./Tech./Santé	89.8%	72.0%	73.4%	2015€

TABLE 2.1 – Statistiques descriptives par domaine académique

Résumé par sexe pour toutes ces années





Sexe	Statistiques			
	 Taux d'insertion	 Emploi cadre	 Emploi stable	 Salaire net
Femmes	89.9%	58.6%	66.5%	1833€
Hommes	90.2%	71.0%	75.6%	2066€
Écart	- 0.3%	-12.4%	-9.1%	-233€

TABLE 2.2 – Statistiques descriptives par sexe académique

Chapitre 3

Analyse Avancée

3.1 Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) est une technique de réduction de dimensionnalité qui sert à explorer et à visualiser les données. Elle est particulièrement utile pour identifier les structures sous-jacentes et pour simplifier la complexité des données en réduisant le nombre de variables. Les composantes principales résultantes sont des combinaisons linéaires des variables initiales et sont sélectionnées de manière à maximiser la variance totale capturée par l'analyse.

Notre recherche s'appuie sur l'ACP pour analyser les données relatives à l'insertion professionnelle des diplômés de l'année 2020. L'objectif est de dégager des insights significatifs sur les facteurs qui influencent les trajectoires professionnelles et les niveaux de rémunération, et ainsi de mieux comprendre la dynamique du marché du travail pour cette promotion particulière.

3.2 Prétraitement des Données et Gestion des Valeurs Manquantes

Une étape préparatoire essentielle à toute analyse en composantes principales est le traitement des valeurs manquantes. Dans notre base de données pour l'année 2020, nous avons identifié une proportion non négligeable de données manquantes qui pourraient biaiser les résultats de notre étude.

Pour contourner ce problème, nous avons mis en œuvre une méthode d'imputation basée sur les forêts aléatoires. Cette technique moderne d'imputation est capable de traiter efficacement les valeurs manquantes en tirant parti des informations disponibles dans l'ensemble de données. Les forêts aléatoires sont particulièrement adaptées à cette tâche en raison de leur capacité à modéliser les interactions complexes et les non-linéarités entre les variables.

Cette approche d'imputation a été appliquée consciencieusement pour rem-

plir les lacunes dans des variables déterminantes telles que le taux d'insertion professionnelle, le taux d'emploi, ainsi que les salaires nets médians. L'objectif principal de cette démarche était de réduire l'impact des valeurs manquantes et de garantir que les résultats de l'ACP soient aussi représentatifs et fiables que possible.

3.3 Préparation des données et exécution de l'ACP

Nous avons réalisé l'ACP en utilisant le package 'FactoMineR' dans R, en spécifiant les variables qualitatives comme variables supplémentaires ('quali.sup') et en appliquant les poids appropriés ('row.w') aux observations. Cette approche pondérée permet une analyse plus représentative des données en prenant en compte la fréquence ou l'importance relative de chaque observation.

Note : Les détails techniques et les scripts R spécifiques à chaque étape du processus ne sont pas inclus dans ce rapport, mais sont documentés séparément pour la transparence et la reproductibilité.

3.4 Résultats de l'Analyse en Composantes Principales

Les résultats de l'ACP sur les données imputées de l'année 2020 révèlent des insights intéressants sur les facteurs influençant l'insertion professionnelle des diplômés.

3.4.1 Cercle des corrélations

À partir de l'ACP, nous avons obtenu un graphique du cercle des corrélations, qui nous montre comment les variables contribuent aux deux premiers axes principaux.

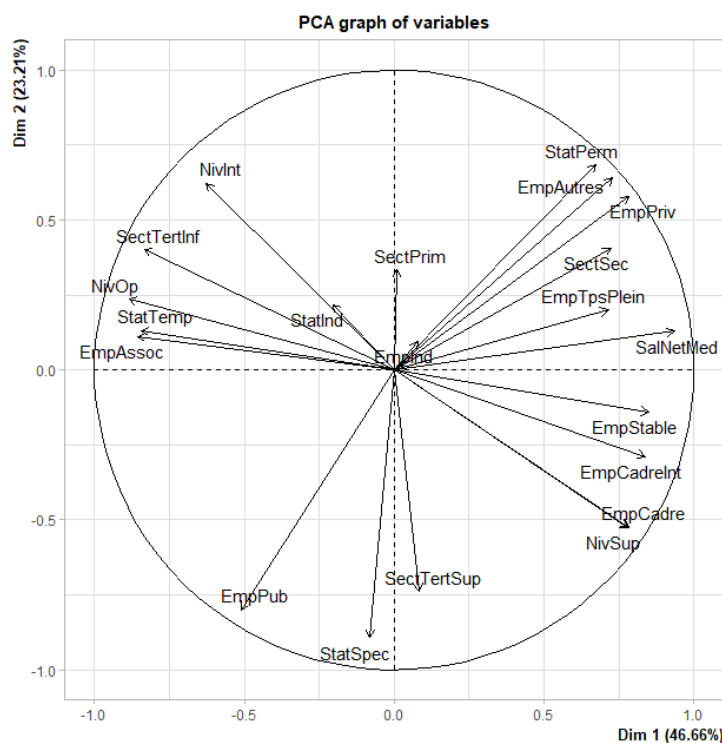


FIGURE 3.1 – Cercle des corrélations des variables de l'ACP.

3.4.2 Interprétation des axes principaux

La **Première Composante Principale** (Axe 1), qui explique 48.16% de la variance totale, est fortement associée à des variables indicatives de la qualité de l'emploi, notamment 'EmpTpsPlein' (emploi à temps plein), 'SalNetMed' (salaire net médian), et 'EmpStable' (emploi stable). Cet axe peut être interprété comme la dimension de la "Stabilité et Rémunération Professionnelles".

La **Deuxième Composante Principale** (Axe 2), expliquant 21.84% de la variance, distingue des caractéristiques telles que 'SectPrim' (secteur primaire) et 'EmpPub' (emploi public), contre 'StatInd' (statut indépendant) et 'NivOp' (niveau opérationnel). Cela pourrait indiquer un clivage entre les types de secteur d'emploi, suggérant ainsi que cette dimension peut être vue comme la "Nature du Secteur d'Emploi".

Ces interprétations nous aident à comprendre que, pour les diplômés de 2020, non seulement la stabilité et la rémunération de leur emploi sont déterminantes, mais aussi le secteur d'activité dans lequel ils s'insèrent.

3.4.3 Interprétation des variables supplémentaires

Dans l'ACP, les variables supplémentaires sont projetées après coup et ne contribuent pas à la détermination des axes. Leur positionnement par rapport aux axes reflète leur association avec les variables actives qui définissent ces axes.

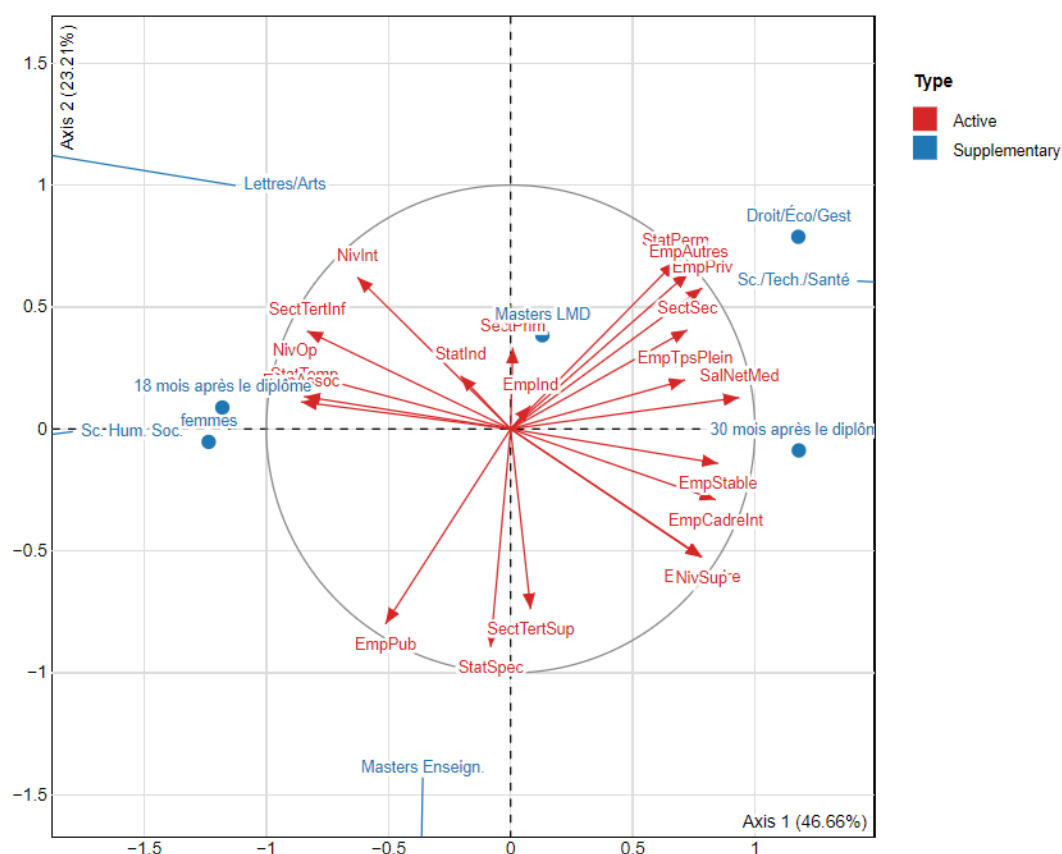


FIGURE 3.2 – Graphique de projection des variables supplémentaires sur les axes de l'ACP.

Comme le montre la Figure 3.2, les modalités des variables supplémentaires se positionnent au barycentre des variables actives les plus fortement associées à elles. Ce positionnement reflète la prédominance de certaines caractéristiques au sein de chaque catégorie. Par exemple, la modalité 'Hommes' se positionne du côté des variables associées à des emplois stables et bien rémunérés, tandis que la modalité 'Femmes' apparaît dans une zone opposée, ce qui pourrait indiquer des différences dans les conditions d'emploi entre les sexes.

Les domaines d'études tels que 'Sc./Tech./Santé' et 'Lettres/Arts' apparaissent à des extrémités opposées sur l'axe des composantes principales. Cela suggère que les caractéristiques d'insertion professionnelle des diplômés de ces

domaines diffèrent significativement. Par exemple, les diplômés en ‘Sc./Tech./Santé’ tendent à être associés à des emplois mieux rémunérés et plus stables, tandis que ceux en ‘Lettres/Arts’ se trouvent dans une zone suggérant des conditions d’emploi potentiellement moins favorables.

La modalité ‘Masters Enseign.’ qui se trouve proche des variables associées à l’emploi public (‘EmpPub’) dans le graphique de l’ACP. Cette proximité montre que les diplômés de cette spécialité ont une tendance marquée à occuper des postes dans le secteur public.

De même, les niveaux de ‘Situ’, qui indiquent le temps écoulé depuis l’obtention du diplôme, montrent que les profils d’insertion peuvent évoluer avec le temps. Les diplômés 30 mois après le diplôme se positionnent différemment par rapport à ceux 18 mois après, ce qui peut indiquer une maturation dans l’insertion professionnelle ou des changements dans les conditions d’emploi avec l’expérience acquise.

Variables supplémentaires qualitatives

Show entries Search:

Variable	Level	Coord	Cos2	V.test	P.value
Genre	Femmes	-1.245	0.950	-224.810	0.000
Genre	Hommes	1.798	0.950	224.810	0.000
Domaine	Sc. Hum. Soc.	-3.753	0.902	-238.560	0.000
Situ	18 mois après le diplôme	-1.123	0.860	-168.820	0.000
Situ	30 mois après le diplôme	1.123	0.860	168.820	0.000
Domaine	Sc./Tech./Santé	2.402	0.742	192.530	0.000
Domaine	Lettres/Arts	-3.874	0.521	-132.270	0.000
Domaine	Droit/Éco/Gest	1.208	0.453	102.330	0.000
Domaine	Masters LMD	0.140	0.054	13.170	0.000
Domaine	Masters Enseign.	-1.831	0.053	-67.170	0.000

Showing 1 to 10 of 10 entries Previous Next

FIGURE 3.3 – Statistiques des variables supplémentaires dans l’ACP.

Le tableau des statistiques associées aux variables supplémentaires, illustré dans la Figure 3.3, fournit des informations supplémentaires sur la significativité de cette projection. Les valeurs élevées des tests de contribution et de significativité (V.test et P.value) confirment que les positions des modalités sur les axes sont statistiquement significatives.

Cette analyse nous permet de comprendre comment les différentes catégories de diplômés sont caractérisées par des ensembles de variables distincts, offrant ainsi des insights précieux sur les tendances d’insertion professionnelle dans notre jeu de données.

3.4.4 Graphique des individus de l'ACP

La représentation des individus dans l'espace des composantes principales donne des indications précieuses sur les similitudes et les différences entre les profils d'insertion professionnelle des diplômés. Chaque point dans le graphique ci-dessous représente un individu, ou plutôt un secteur disciplinaire qui constitue des groupes d'individus, avec sa position reflétant ses scores sur les deux premières composantes principales.

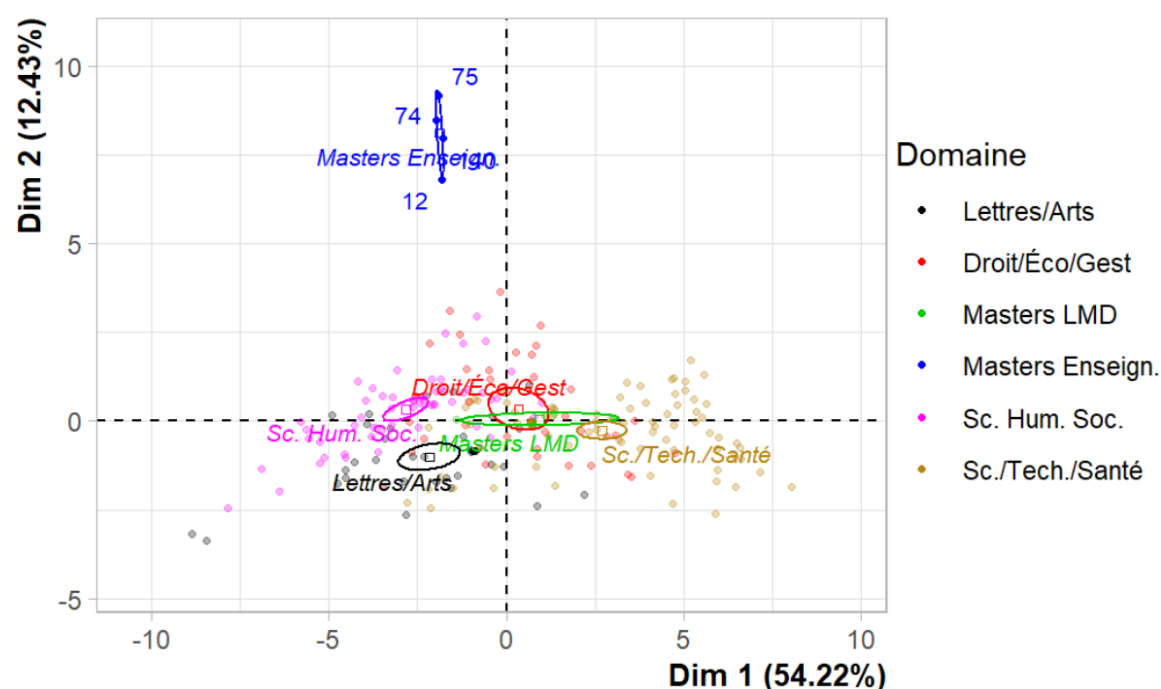


FIGURE 3.4 – Graphique des individus.

Comme montre dans la Figure 3.4, certains groupes de diplômés, tels que ceux avec un Master en Enseignement, se distinguent nettement des autres et sont localisés dans une région spécifique de l'espace de l'ACP. Cette localisation suggère des caractéristiques uniques en termes de type d'emploi ou de conditions de travail associées à ces diplômés.

L'analyse montre également une dispersion considérable parmi les individus, indiquant une variabilité dans l'insertion professionnelle au sein même des secteurs disciplinaires d'études. Les diplômés de certains domaines, ont des profils bien diversifiés

3.5 Résultats du clustering

La classification hiérarchique a permis de révéler des groupes distincts au sein des diplômés de 2020, mettant en lumière les divers chemins d'insertion

professionnelle.

L'analyse a révélé la présence de quatre groupes distincts au sein de la population étudiée, comme illustré dans le dendrogramme (Figure 3.6). Ces clusters ont été interprétés en fonction des caractéristiques dominantes des individus qui les composent, observées dans le plan factoriel (Figure 3.5).

- Le **cluster 1** est composé principalement de diplômés 18 mois après l'obtention du diplôme, avec une forte représentation des femmes et des domaines des sciences humaines et sociales.
- Le **cluster 2** est marqué par une forte présence de diplômés du domaine de l'Enseignement, ayant un taux d'insertion professionnelle élevé et des emplois stables, dans le secteur public.
- Le **cluster 3** se distingue par une concentration d'individus issus des domaines scientifiques, technologiques et de santé, ainsi que de l'économie et la gestion, qui ont une insertion dans des secteurs à forte demande et potentiellement de meilleures conditions d'emploi.
- Le **cluster 4** regroupe des individus qui sont plus avancés dans leur parcours post-diplôme (30 mois après), indiquant une phase de stabilisation dans leur carrière avec des emplois plus stables et établis.

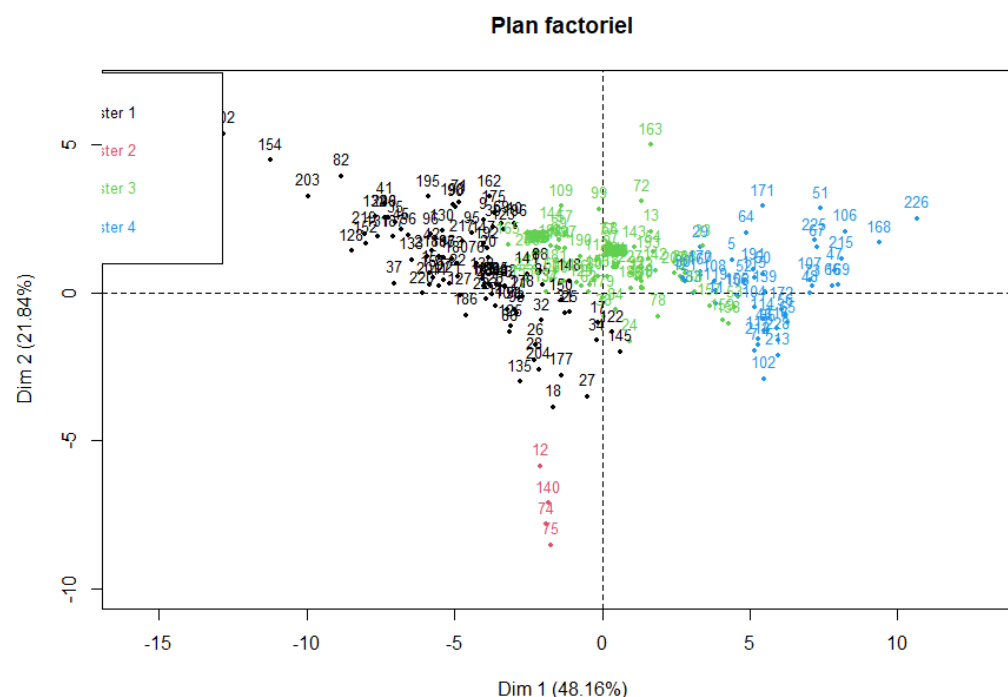


FIGURE 3.5 – Représentation des individus sur le plan factoriel avec la segmentation en clusters.

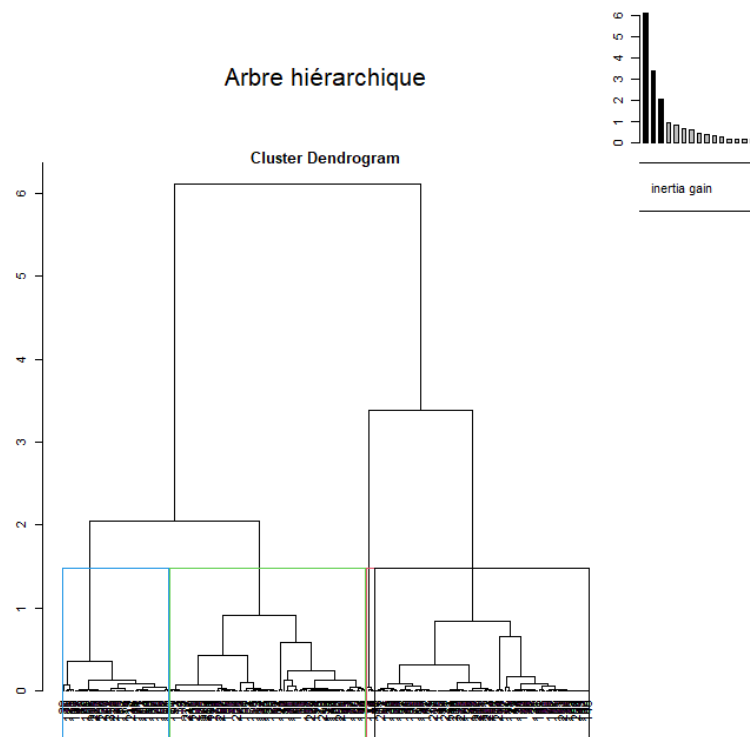


FIGURE 3.6 – Dendrogramme résultant du clustering hiérarchique.

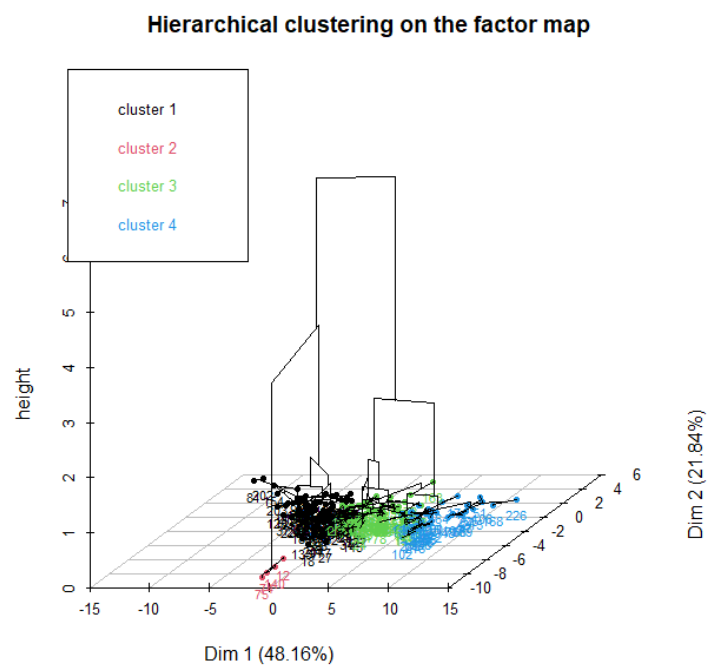


FIGURE 3.7 – Dendrogramme 3D résultant du clustering hiérarchique.

3.6 Analyse de Régression

Nous avons examiné les facteurs influençant le salaire net médian des diplômés à travers une régression linéaire pondérée, en considérant des variables telles que le genre et les catégories de statuts d'emploi. Les résultats mettent en évidence les relations entre ces facteurs et le salaire net médian, en prenant en compte les différences de genre et la nature du contrat de travail.

3.6.1 Résultats du modèle de régression

Le tableau ci-dessous présente les estimations des coefficients du modèle de régression linéaire pondérée, indiquant l'influence de chaque prédicteur sur le log du salaire net médian. Les coefficients significatifs suggèrent des différences systématiques dans les salaires qui peuvent être attribuées à des facteurs tels que le genre et le statut d'emploi.

TABLE 3.1 – Régression linéaire pondérée

	<i>Dependent variable :</i>
	log(SalNetMed)
GenreHommes	0.049*** (0.005)
StatPerm	0.002*** (0.0002)
StatTemp	-0.004*** (0.0002)
StatInd	-0.006*** (0.001)
Constant	7.589*** (0.015)
Observations	228
R ²	0.864
Adjusted R ²	0.862
Residual Std. Error	1.155 (df = 223)
F Statistic	355.654*** (df = 4; 223)
<i>Note :</i> *p<0.1 ; **p<0.05 ; ***p<0.01	

Les hommes gagnent en moyenne significativement plus que les femmes, avec un coefficient de 0.0492615 ($p \leq 2e-16$), confirmant l'existence d'une disparité salariale entre les sexes dans le marché du travail actuel. Cette constatation est cohérente avec les études antérieures qui ont identifié le genre comme un déterminant clé des différences de salaire.

3.6.2 Interprétation des effets des variables

La Figure 3.8 représente graphiquement l'effet des variables sélectionnées sur le salaire net médian. L'impact du genre est particulièrement prononcé, comme on vient de le voir dans les autres analyses.

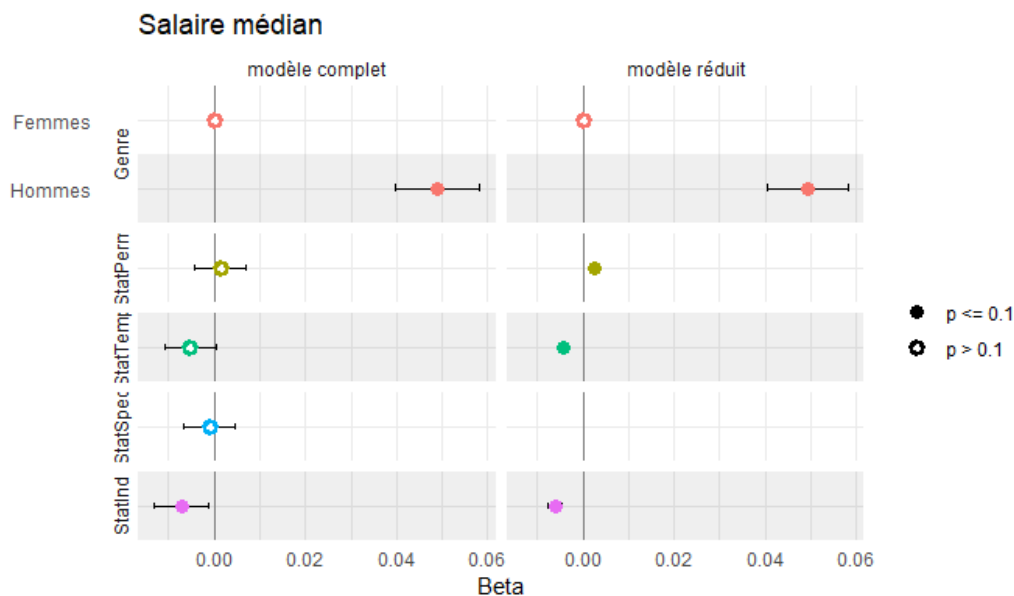


FIGURE 3.8 – Effets estimés des variables explicatives sur le salaire net médian pour les modèles complet et réduit.

Suite à une procédure de sélection de modèle basée sur le critère AIC pour optimiser la simplicité et l'efficacité explicative du modèle, la variable représentant le statut spécialisé a été exclue. Le modèle réduit, inclue le genre, le statut d'emploi permanent, temporaire, et indépendant, a conservé une forte capacité explicative avec un R^2 ajusté de 86.21%. Ce modèle confirme l'effet significatif du genre sur le salaire net médian, avec un coefficient pour les hommes de 0.0492615 ($p \leq 2e-16$), et montre l'importance des statuts d'emploi permanent et temporaire dans la détermination des salaires.

3.6.3 Analyse des effets des variables sur le salaire net médian

L'analyse des effets des variables sur le log du salaire net médian met en lumière les différences systématiques attribuables aux caractéristiques démographiques et contractuelles des diplômés. Cette analyse, présentée dans la figure ci-dessous, fournit une estimation quantifiée de l'influence de chaque variable, accompagnée de mesures d'incertitude statistique sous forme d'intervalles de confiance.

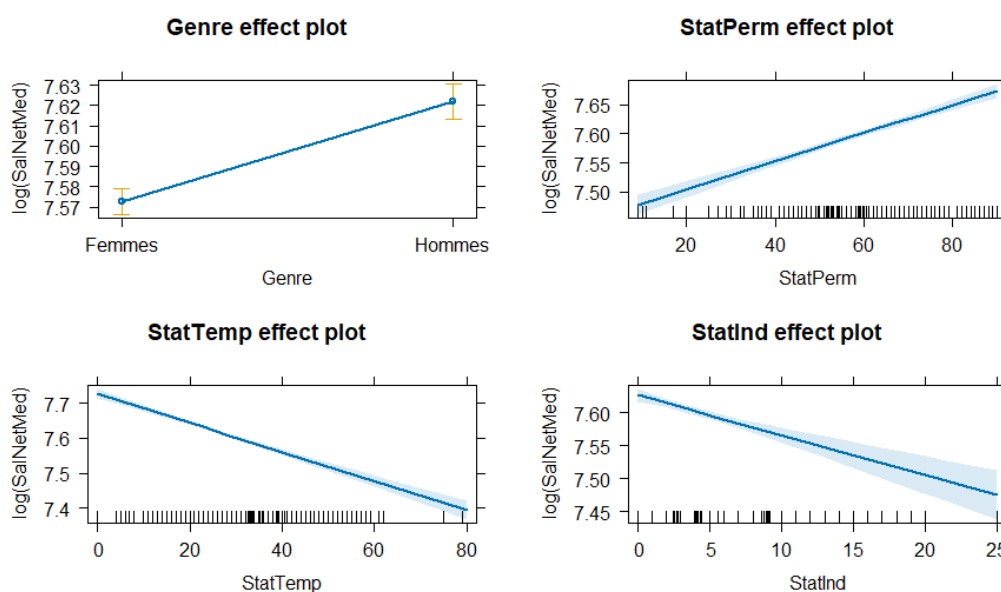


FIGURE 3.9 – Effets estimés des variables explicatives sur le log du salaire net médian. Les graphiques détaillent les contributions marginales des variables de genre, de statut d'emploi permanent, de statut d'emploi temporaire et de statut d'emploi indépendant, révélant les tendances et les significativités associées.

Le graphique dédié à l'effet du genre confirme la présence d'une disparité salariale marquée, corroborant les conclusions d'études antérieures sur l'écart de rémunération entre les sexes dans le marché du travail post-universitaire. Concrètement, l'effet positif significatif associé au genre masculin suggère une prime salariale pour les hommes par rapport aux femmes, ce qui mérite une attention accrue dans les discussions sur l'équité salariale.

En ce qui concerne les variables liées au statut d'emploi, les effets varient, reflétant la complexité des trajectoires professionnelles et l'interaction avec d'autres facteurs socio-économiques. Le statut d'emploi permanent est comme un facteur positif, indiquant une valorisation sur le marché du salaire pour les contrats à long terme. À l'inverse, le statut d'emploi temporaire et indépendant est associé à des salaires inférieurs, mettant en évidence l'importance des conditions de travail sécurisées et stables pour assurer une rémunération adéquate.

3.7 Vérification des hypothèses de la régression linéaire

La validité des conclusions tirées à partir d'un modèle de régression linéaire repose sur la satisfaction de plusieurs hypothèses clés. Il est essentiel de vérifier ces hypothèses pour s'assurer de la fiabilité des estimations des paramètres et des tests d'hypothèses associés. Les hypothèses suivantes ont été examinées pour le modèle de régression employé dans cette étude :

3.7.1 Diagnostic du modèle de régression

Un élément important de l'analyse de régression est la vérification que le modèle satisfait les hypothèses de la régression linéaire. La figure ci-dessous montre le graphique des résidus par rapport aux valeurs ajustées, qui est utilisé pour évaluer l'homoscédasticité et la linéarité des résidus.

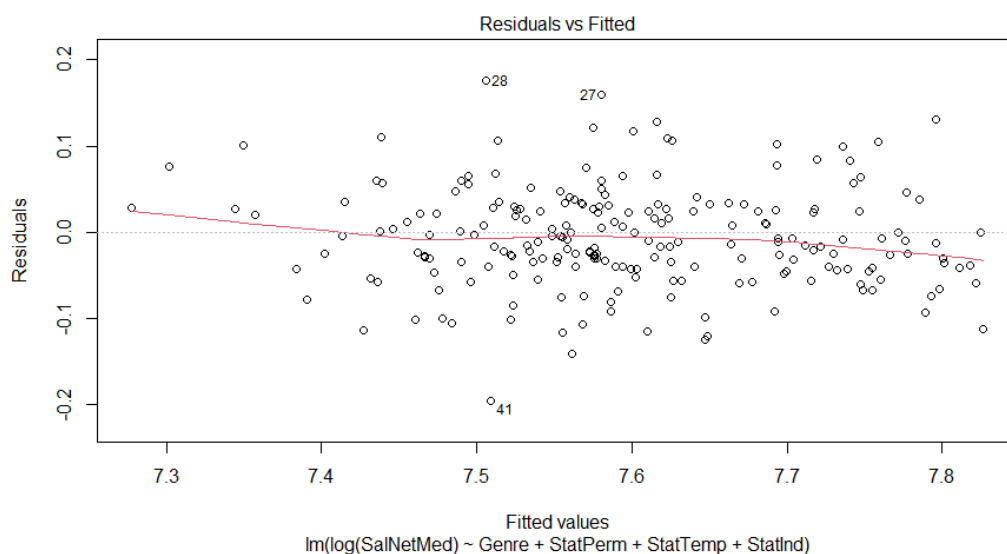


FIGURE 3.10 – Graphique des résidus par rapport aux valeurs ajustées pour le modèle de régression linéaire.

La distribution apparemment aléatoire des résidus autour de la ligne horizontale dans la Figure 3.10 montre l'absence de structures ou de tendances systématiques, ce qui est un indicateur que la relation linéaire spécifiée dans le modèle peut être appropriée.

En complément du diagnostic graphique, le test Rainbow a été effectué pour évaluer la linéarité de la relation entre les variables indépendantes et dépendantes. Le test Rainbow produit une statistique de test de 1.3265 avec

une valeur p de 0.06909, ce qui ne permet pas de rejeter l'hypothèse nulle de linéarité à un seuil de signification de 5%.

3.7.2 Vérification de l'indépendance des résidus

Pour confirmer l'hypothèse d'indépendance des résidus, un diagnostic dans l'analyse de régression, nous avons étudié l'autocorrélation des erreurs à l'aide de l'autocorrélation des résidus et du test de Ljung-Box.

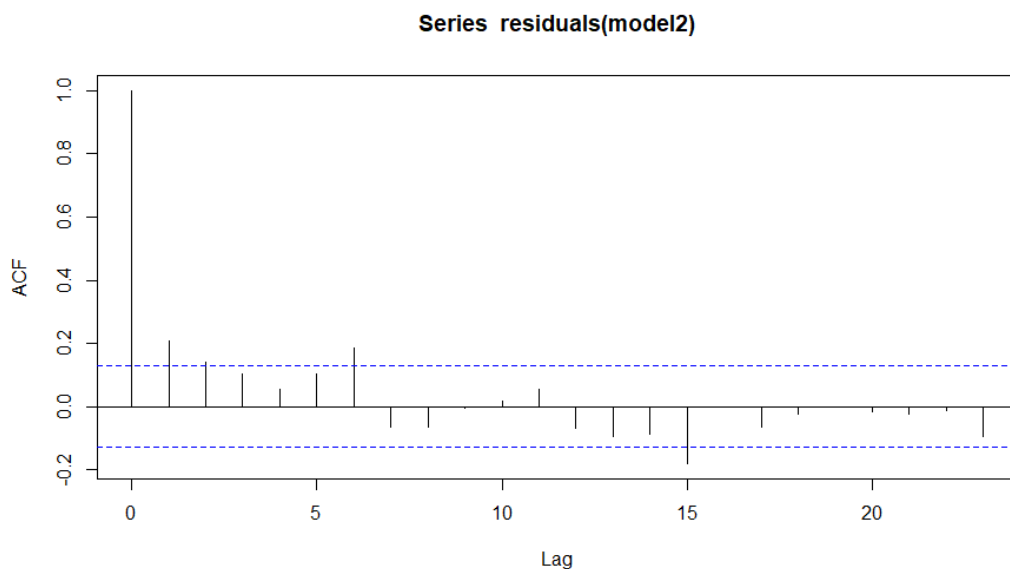


FIGURE 3.11 – Fonction d'autocorrélation des résidus du modèle de régression linéaire ajusté.

La Figure 3.11 montre la fonction d'autocorrélation pour les résidus du modèle. Les barres se situent majoritairement dans les limites de confiance, ce qui montre l'absence d'autocorrélation. Cependant, le test de Ljung-Box a donné un résultat de $X^2 = 10.134$ avec une valeur p de 0.001455, indiquant une autocorrélation significative au niveau des résidus.

Ce résultat montre que malgré l'apparence d'indépendance dans la fonction d'autocorrélation, il existe des preuves statistiques d'autocorrélation résiduelle, ce qui pourrait indiquer des lacunes dans le modèle actuel ou la présence de variables omises qui influencent les résidus. Il est essentiel d'aborder cette question pour améliorer la fiabilité du modèle de régression.

3.7.3 Test d'hétéroscédasticité

Pour valider l'hypothèse d'homogénéité des variances des erreurs (homoscédasticité) dans notre modèle de régression linéaire, nous avons effectué le test de Breusch-

Pagan. Ce test est important pour évaluer la fiabilité des estimations des erreurs standards des coefficients du modèle.

Les résultats du test de Breusch-Pagan montre fortement la présence d'hétéroscédasticité dans les résidus du modèle. Avec une statistique de test de 218127 et une valeur-p pratiquement nulle, nous rejetons l'hypothèse d'homoscédasticité.

$$BP = 218127, \quad df = 4, \quad p\text{-value} < 2.2 \times 10^{-16}$$

Ces résultats indiquent que la variance des erreurs n'est pas constante à travers les niveaux des prédicteurs. L'hétéroscédasticité entraînent des estimations biaisées des erreurs standards, ce qui compromet l'exactitude des tests d'hypothèses et des intervalles de confiance. Des méthodes correctives, telles que l'utilisation de robustes erreurs standards ou la transformation des variables, seront envisagées pour remédier à ce problème et garantir des inférences plus fiables à partir du modèle de régression.

Le diagnostic des observations influentes pour assurer la robustesse des modèles de régression. Le graphique d'influence présenté ci-dessous montre les résidus standardisés en fonction des valeurs de levier (hat values), avec la taille des cercles représentant la distance de Cook pour chaque observation.

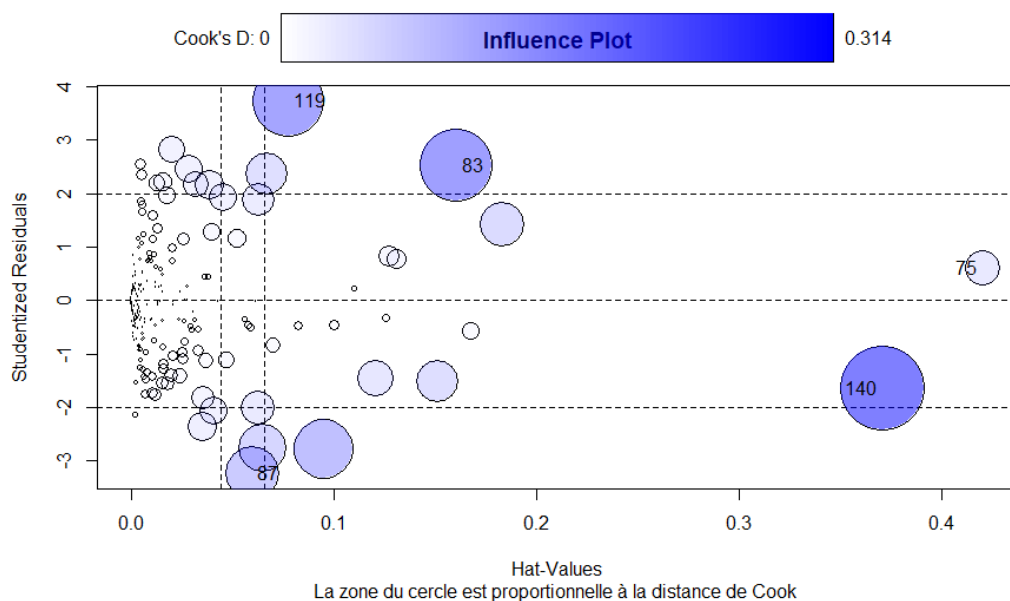


FIGURE 3.12 – Graphique d'influence indiquant les observations potentiellement influentes. Les cercles représentent la distance de Cook, avec un intérêt particulier pour les observations dont la distance de Cook dépasse 1, ce qui indique une influence substantielle sur le modèle.

Comme indiqué sur le graphique, certaines observations présentent une distance de Cook nettement supérieure à 1, ce qui suggère qu'elles ont une

influence disproportionnée sur les estimations des paramètres du modèle. Ces points sont des candidats potentiels pour une analyse plus approfondie, car leur présence peut fausser les résultats de la régression et potentiellement conduire à des conclusions erronées. Un examen plus minutieux de ces points est de les exclusion du modèle ou l'investigation des raisons de leur influence excessive (méthode que nous ferons par la suite).

3.7.4 Normalité des résidus

L'analyse de la normalité des résidus est essentielle dans les modèles de régression linéaire, car elle conditionne la validité des tests statistiques utilisés pour l'inférence. Pour évaluer cette normalité, nous utilisons un graphique quantile-quantile (Q-Q plot) des résidus standardisés issus de notre modèle.

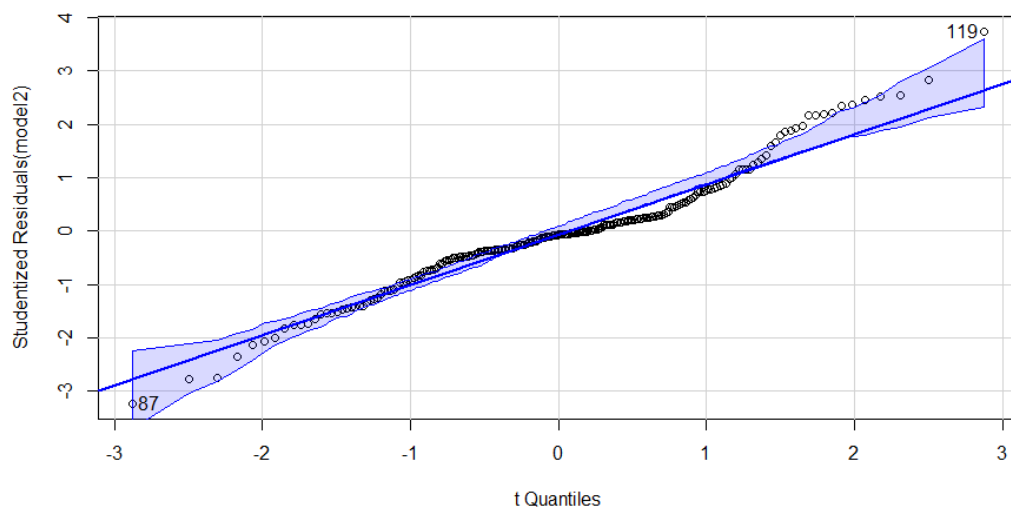


FIGURE 3.13 – Graphique Q-Q des résidus standardisés du modèle. Ce graphique compare la distribution des résidus avec une distribution normale théorique. Les points s'écartant de la ligne représentent des déviations par rapport à la normalité.

En parallèle, le test de Shapiro-Wilk a été réalisé pour fournir une évaluation statistique de la normalité. Avec une valeur-p de 0.06333, le test ne permet pas de rejeter l'hypothèse de normalité au seuil de 5%. Cependant, cette valeur est assez proche du seuil critique, la normalité des résidus pourrait être une préoccupation dans des échantillons plus grands ou avec des tests plus sensibles.

3.7.5 Analyse de la multicolinéarité

La multicolinéarité entre les prédicteurs d'un modèle de régression peut compromettre l'interprétation des coefficients de régression et réduire la précision des estimations. Pour évaluer la présence de multicolinéarité, nous avons calculé le facteur d'inflation de la variance (VIF) pour chaque variable explicative du modèle.

Les résultats des VIF sont les suivants :

- Genre : 1.147
- Statut d'emploi permanent (StatPerm) : 1.401
- Statut d'emploi temporaire (StatTemp) : 1.433
- Statut d'emploi indépendant (StatInd) : 1.021

Ces valeurs suggèrent l'absence de multicolinéarité importante dans le modèle, car toutes sont nettement inférieures au seuil communément admis de 5 . Cela indique que chaque prédicteur apporte des informations uniques qui contribuent à la prédiction du log du salaire net médian, renforçant ainsi la fiabilité des estimations des coefficients obtenus dans le modèle de régression linéaire ajusté.

3.7.6 Gestion des observations influentes

La présence d'observations influentes, révélée par des distances de Cook supérieures à un seuil critique, nécessite une attention particulière en raison de leur potentiel de biais significatif dans les estimations des paramètres du modèle. Une méthode courante pour atténuer ce problème consiste à retirer ces observations de l'ensemble de données et à recalculer les estimations du modèle.

Dans notre étude, nous avons identifié un ensemble d'observations dont les résidus standardisés suggèrent une influence disproportionnée sur le modèle de régression linéaire. Ces points ont été exclus de l'analyse pour améliorer la robustesse et la fiabilité de notre modèle. Ce processus de nettoyage des données est une étape préliminaire standard avant de procéder à une ré-estimation des paramètres du modèle.

Après l'exclusion des observations influentes, une nouvelle base de données a été constituée et a servi à recalculer le modèle de régression. Les hypothèses du modèle ont été réexaminées, y compris l'homoscédasticité des erreurs, la normalité des résidus et l'indépendance des observations. Les résultats de ces diagnostics post-épuration sont présentés dans les sections suivantes, confirmant la validité des ajustements effectués et la robustesse du modèle révisé.

3.7.7 Comparaison des modèles avant et après correction des valeurs influentes

L'analyse comparative des modèles de régression avant et après correction des valeurs influentes révèle des améliorations notables dans la précision des estimations des coefficients et la qualité globale du modèle.

Le premier modèle, avant correction, présentait déjà un R-carré ajusté élevé de 0.8621, suggérant une bonne adéquation du modèle avec les données. Cependant, la présence d'observations influentes pouvait biaiser les estimations des paramètres. Le second modèle, après correction, montre un R-carré ajusté encore plus élevé de 0.9655, indiquant une explication substantiellement améliorée de la variabilité du salaire net médian par les variables choisies.

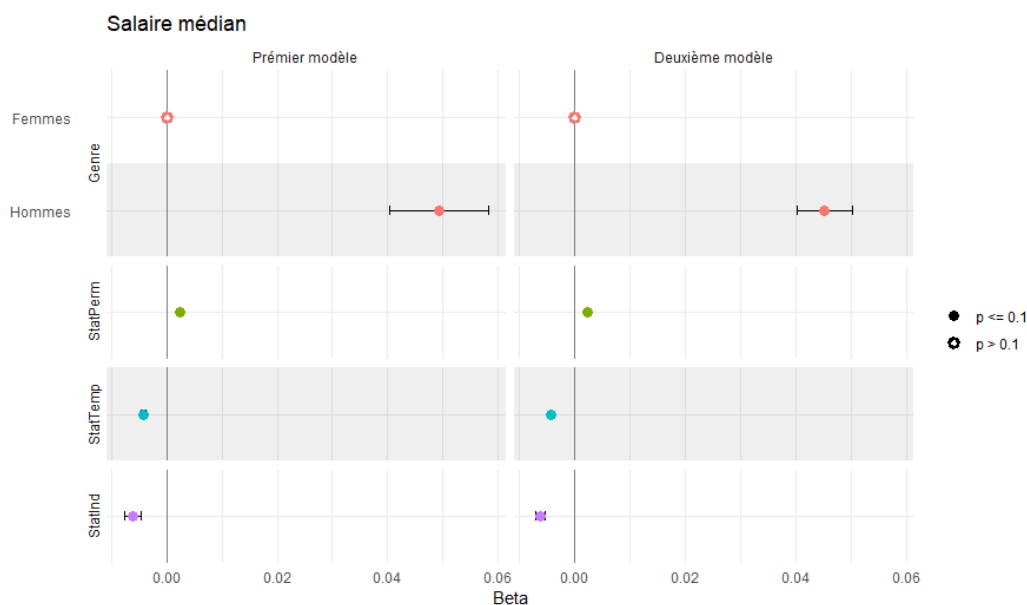


FIGURE 3.14 – Comparaison des coefficients estimés du modèle de régression avant et après correction des valeurs influentes. Les intervalles de confiance à 90% sont présentés pour chaque coefficient dans les deux modèles, illustrant la stabilité des estimations après correction.

Les coefficients de 'GenreHommes' et les variables associées au statut d'emploi dans les deux modèles sont statistiquement significatifs et conservent leur signe, mais l'ampleur de leur influence varie entre les modèles. Notamment, la variable 'Genre = Hommes' a un coefficient légèrement plus élevé après correction, renforçant la preuve de l'écart de rémunération basé sur le genre.

La réduction de l'erreur standard des résidus dans le deuxième modèle indique que les prédictions sont plus précises, et la baisse significative de l'erreur résiduelle standard suggère que les résidus sont plus serrés autour de 0, ce qui implique une meilleure adéquation du modèle.

3.7.8 Analyse des hypothèses statistiques après correction

La correction des observations influentes a permis une réestimation du modèle de régression. Les tests statistiques suivants ont été appliqués pour vérifier la validité des hypothèses du modèle révisé.

Test rainbow pour la linéarité

Le test Rainbow a donné une valeur de 0.93865 avec une valeur-p de 0.6102, ne fournissant aucune preuve contre l'hypothèse de linéarité du modèle. Ceci suggère que la relation entre les variables indépendantes et la variable dépendante est bien modélisée par une fonction linéaire.

Test de Ljung-Box pour l'indépendance des résidus

Le test de Ljung-Box a produit une statistique de test de 2.2126 et une valeur-p de 0.1369, indiquant l'absence d'autocorrélation significative dans les résidus du modèle. Cela confirme que les résidus sont indépendants les uns des autres, une condition nécessaire pour les inférences valides dans le modèle de régression linéaire.

Test de Shapiro-Wilk pour la normalité des résidus

Le test de Shapiro-Wilk a donné un résultat de $W = 0.97906$ avec une valeur-p de 0.01623, indiquant une déviation de la normalité des résidus à un niveau de signification de 5%. Bien que cette valeur-p suggère que la distribution des résidus n'est pas parfaitement normale, le degré de déviation n'est pas extrême. Cela peut nécessiter une enquête supplémentaire ou l'utilisation de techniques robustes pour assurer la validité des tests statistiques.

Facteurs d'inflation de la variance

Les valeurs VIF pour chaque prédicteur sont bien en dessous du seuil communément admis de 5 ou 10

3.8 Modélisation par Arbre de Régression

Après avoir examiné les relations linéaires entre les variables explicatives et le log du salaire net médian à travers des modèles de régression linéaire, nous explorons maintenant des méthodes non linéaires pour capturer des relations plus complexes et des interactions potentielles entre les prédicteurs. À cet égard, les arbres de régression offrent une approche flexible et interprétable

pour modéliser des structures de données intrinsèquement non linéaires et hiérarchiques.

Les arbres de régression sont des outils de modélisation prédictive qui segmentent l'espace des prédicteurs en un ensemble de régions simples. En utilisant la stratégie de division récursive binaire, l'arbre de régression cherche à identifier les points de scission qui maximisent la différence de la variable de réponse entre les branches. Cela permet non seulement d'isoler les effets des variables dans des sous-groupes spécifiques mais aussi d'illustrer l'importance relative des différentes variables dans la prédiction de la réponse.

Dans la section suivante, nous construirons un arbre de régression en utilisant la variable de réponse transformée et évaluons sa performance en tant que modèle prédictif comparativement aux analyses précédentes. Ce processus comprendra la création de nouvelles variables si nécessaire, l'ajustement de l'arbre, l'élagage pour éviter le surajustement et l'interprétation des résultats obtenus.

3.8.1 Préparation des données pour l'arbre de régression

Dans l'étape préalable de notre analyse, nous avons utilisé l'Analyse en Composantes Principales (ACP) pour réduire la dimensionnalité de notre ensemble de données et pour identifier les variables les plus informatives concernant la qualité de l'emploi. La première composante principale, qui explique 48.16% de la variance totale, s'est avérée être fortement liée à des indicateurs clés de la qualité de l'emploi, notamment l'emploi à temps plein ('EmpTps-Plein'), le salaire net médian ('SalNetMed'), et la stabilité de l'emploi ('EmpStable'). Cette composante peut donc être interprétée comme représentant la dimension de la "Stabilité et Rémunération Professionnelles".

En se basant sur cette interprétation, nous avons créé une nouvelle variable catégorielle à partir des scores des individus sur cette composante principale. Les individus ont été classés en quatre groupes distincts selon les quartiles des scores : 'Très faible', 'Faible', 'Moyen', 'Élevé'. Cette classification nous permet d'intégrer une mesure synthétique de la qualité de l'emploi dans notre modèle d'arbre de régression.

L'usage de cette nouvelle variable catégorielle est double : d'une part, elle résume plusieurs aspects importants de la qualité de l'emploi en une seule mesure, et d'autre part, elle nous permet d'explorer les non-linéarités et les interactions potentielles entre la qualité de l'emploi et le salaire net médian. Dans la section suivante, nous utiliserons cette variable catégorielle comme prédicteur dans notre arbre de régression pour évaluer son influence sur le salaire net médian, en la comparant aux autres prédicteurs du modèle.

3.8.2 Construction de l'Arbre de Régression

L'arbre de régression est un outil puissant pour visualiser et interpréter les interactions complexes entre les variables explicatives et la variable de réponse. Dans notre analyse, nous avons catégorisé le salaire net médian en trois groupes : inférieur à 2000 euros, entre 2000 et 2500 euros, et supérieur à 2500 euros. Cette catégorisation est basée sur des seuils pertinents identifiés dans la distribution des salaires.

Nous avons ensuite construit un arbre de régression en utilisant le genre, le domaine d'études et les catégories de la première composante principale comme prédicteurs. Le paramètre de complexité (*cp*) a été fixé à une valeur très faible pour permettre à l'arbre de croître pleinement, et le nombre minimum d'observations requises pour diviser un nœud (*minspl*) a été fixé à 2.

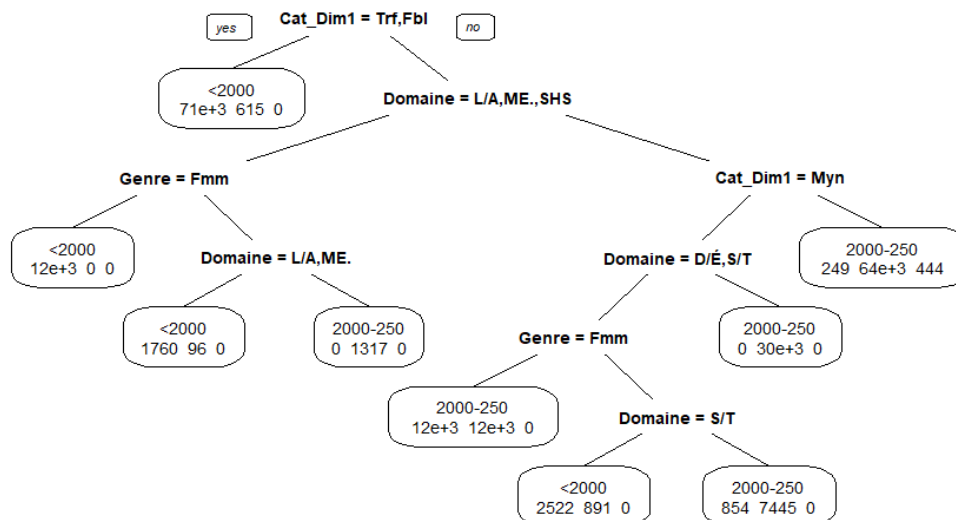


FIGURE 3.15 – Arbre de régression illustrant la relation entre les variables catégorielles et le salaire net médian. Chaque nœud de l'arbre représente une scission basée sur une variable, divisant les observations en groupes homogènes en termes de salaire.

La visualisation de l'arbre, illustrée dans la Figure 3.15, montre les différentes scissions et les critères utilisés par l'algorithme pour segmenter la population étudiée. Par exemple, la première division est basée sur la première composante principale, ce qui reflète l'importance de la "Stabilité et Rémunération Professionnelles" dans la détermination des catégories salariales. Les divisions subséquentes affinent ces groupes en fonction du genre et du domaine d'études, ce qui révèle des tendances intéressantes dans les données.

Cette approche a permis de mettre en évidence des tendances claires, telles que la différenciation par genre et domaine d'études dans les groupes salariaux.

inférieurs et moyens. L'arbre de régression confirme et quantifie l'influence de ces variables sur le salaire net médian, tout en offrant une représentation intuitive de la structure des données.

Chapitre 4

Conclusion

Après une analyse minutieuse et approfondie des facteurs influençant l’insertion professionnelle des diplômés de Master, cette étude met en lumière des dynamiques complexes et des disparités notables sur le marché du travail. Le recours à des méthodes statistiques avancées, incluant la régression linéaire pondérée, l’Analyse en Composantes Principales (ACP), ainsi que les arbres de régression, a permis de déchiffrer les influences du genre, du domaine d’études, et de la qualité de l’emploi sur le salaire net médian.

Le genre s’est révélé être un prédicteur significatif, avec les hommes gagnant systématiquement plus que les femmes, ce qui reflète et confirme la persistance de l’écart salarial de genre dans le contexte actuel. Cette étude rejoint ainsi le corpus de recherches existantes qui témoignent des inégalités de rémunération sur le marché du travail, soulignant la nécessité d’efforts continus et ciblés pour parvenir à une égalité salariale.

L’analyse a également démontré l’importance de la stabilité de l’emploi et du domaine d’études. Les diplômés des domaines des Sciences/Technologie/Santé et Droit/Économie/Gestion tendent à jouir d’une meilleure insertion professionnelle, contrastant avec les domaines des Lettres et des Arts, laissant croire une adéquation plus forte entre les compétences acquises dans ces domaines et les demandes du marché du travail. Ces domaines semblent mieux préparer les étudiants aux postes bien rémunérés, probablement en raison de la demande soutenue pour des compétences spécialisées.

En dépit de la performance des modèles de régression, l’analyse a révélé la présence de valeurs influentes et d’hétéroscédasticité dans les données, qui ont été adressées par des techniques statistiques correctives. L’exclusion des observations influentes et l’ajustement des modèles ont permis d’obtenir des résultats plus fiables et robustes.

Enfin, l’arbre de régression a offert une illustration visuelle des relations non linéaires et des interactions entre les prédicteurs. Cette modélisation a renforcé la compréhension de la façon dont différents facteurs, y compris la qualité de l’emploi mesurée par la première composante principale, interagissent pour influencer le salaire net médian.

Cette recherche contribue significativement à la littérature sur l'insertion professionnelle des diplômés de Master en France. Elle fournit des insights précieux pour les décideurs politiques et les institutions éducatives, en mettant en évidence les axes sur lesquels concentrer les efforts pour améliorer l'insertion professionnelle et réduire les inégalités salariales. De plus, elle souligne l'importance de poursuivre la recherche dans ce domaine, en adoptant des méthodes analytiques sophistiquées pour comprendre les facteurs sous-jacents qui influencent les trajectoires professionnelles des jeunes diplômés.

Les pistes pour les recherches futures incluent :

1. **Études Longitudinales** : Suivre les cohortes de diplômés sur une période plus étendue permettrait de saisir l'évolution de l'insertion professionnelle et des écarts salariaux sur le long terme. Les transitions entre différents types d'emplois, les progressions de carrière et les changements de salaires apporteraient une compréhension plus nuancée des dynamiques de travail.
2. Une **Analyse Segmentée** par Domaine et Genre : Une analyse plus détaillée, centrée sur les intersections spécifiques entre genre et domaine d'études, pourrait révéler des disparités plus subtiles et fournir des insights pour des politiques d'égalité plus ciblées.
3. Une évaluation de l'**Impact des Politiques d'Éducation et d'Emploi** : Évaluer l'efficacité des politiques publiques actuelles en matière d'éducation et d'insertion professionnelle pourrait mener à des recommandations pour des ajustements stratégiques, visant à une meilleure adéquation entre formation et emploi.
4. Une investigation de l'**Influence des Changements Économiques et Technologiques** : Avec l'évolution rapide du paysage économique et technologique, il est crucial d'examiner comment ces changements affectent l'insertion professionnelle. Des études futures pourraient intégrer l'impact de l'automatisation, de la numérisation et des nouvelles formes d'organisation du travail.
5. Des **Études Comparatives Internationales** : Une comparaison avec des données issues d'autres contextes nationaux enrichirait la compréhension des facteurs globaux et locaux qui influencent l'insertion professionnelle.
6. L'incorporation d'**approches Qualitatives** : Compléter les analyses quantitatives par des méthodes qualitatives permettrait de saisir les expériences vécues des diplômés, leurs perceptions du marché du travail et les stratégies qu'ils déploient pour naviguer dans leur carrière.
7. Une exploration de l'importance du **développement Professionnel et Formation Continue** : Explorer le rôle de la formation continue et du développement professionnel dans l'augmentation de la valeur sur le marché du travail des diplômés.

8. Une analyse de l'**Impact de la Politique de Diversité et d'Inclusion** : Analyser comment les politiques d'inclusion en entreprise et dans l'enseignement supérieur affectent les trajectoires professionnelles des groupes sous-représentés.

Cette recherche éclaire les politiques actuelles et futures concernant l'enseignement supérieur et l'insertion professionnelle. Les résultats invitent à une réflexion continue sur les stratégies pour promouvoir l'égalité des chances et répondre aux besoins changeants des diplômés et du marché du travail. Les perspectives futures ouvrent des voies passionnantes pour des contributions significatives dans le domaine de l'emploi et de l'éducation.

Bibliographie

- COUPPIÉ, Thomas, et Dominique EPIPHANE. “Et les femmes devinrent plus diplômées que les hommes...” *Céreq Bref*, numéro 373, 2019, pages 1-4.
- . “La relation genre-insertion at-elle évolué en 20 ans?” *Essentiels*, numéro 1, 2018, pages 141-49.
- DIRECTION DE L’ÉVALUATION, DE LA PROSPECTIVE ET DE LA PERFORMANCE (DEPP). Repères et Références statistiques. 2021, Disponible sur : <https://www.education.gouv.fr/direction-de-l-evaluation-de-la-prospective-et-de-la-performance-depp-12389>.
- FACK, Gabrielle, et Élise HUILLERY. “Enseignement supérieur : pour un investissement plus juste et plus efficace”. *Notes du conseil danalyse economique*, tome 68, numéro 8, 2021, pages 1-12.
- GEORGES-KOT, Simon. “Écarts de rémunération femmes-hommes : surtout l’effet du temps de travail et de l’emploi occupé”. *Insee première*, tome 1803, 2020, pages 1-4.
- MÉNARD, Boris. “Quel effet de la série et de la mention du baccalauréat sur l’insertion des diplômés de master?” *ÉCHANGES*, page 29.
- MEURS, Dominique, et Sophie PONTHEUX. “Une mesure de la discrimination dans l’écart de salaire entre hommes et femmes”. *Économie et statistique*, tome 337, numéro 1, 2000, pages 135-58.
- MINISTÈRE DE L’ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L’INNOVATION (MENESR-DGESIP/DGRI-SIES). “Enquête 2015 sur l’insertion professionnelle des diplômés de l’université”, 2015, 1 rue Descartes, 75231 Paris Cédex 05, France, www.enseignementsup-recherche.gouv.fr/pid24624/taux-insertion-professionnelle-des-diplomes-universite.html.

sample

Appendices

.1 Dictionnaire des variables